

Leveraging collaborative knowledge base of Wikipedia for Text Analytics

Mohamed Minhaj

Associate Professor - Systems

SDMIMD, Mysuru

mminhaj@sdmimd.ac.in



Shri Dharmasthala Manjunatheshwara

Research Center for Management Studies (SDM RCMS)

SDM Institute for Management Development (SDMIMD)

Mysuru - 570 011

(C) Applied Research Project 2020, SDM Research Center for Management Studies (SDM RCMS), SDMIMD, Mysuru

Note:

All views expressed in this work are that of the author(s). SDM RCMS does not take any responsibility for the views expressed herein by the author(s).

No part of this publication can be reproduced or transmitted in any form or by any means, without prior permission of the publisher.

Preface

SDM Research Centre for Management Studies (SDM RCMS), since inception, has endeavoured to promote research in the field of business management in different ways. One of the initiatives undertaken, in this direction, was to enable Faculty to pursue applied research projects in the realm of management.

Broadly, an applied research project is expected to answer a specific question, determine why something is failed or succeeded, solve a specific and pragmatic problem relating to the business, economy, or policies. It may also intend to study the relationship and applicability of management theories or principles to the solution of a problem. It uses the data on a specific set of circumstances directly for real world application with the goal of relating the results to a situation. Hence an applied research is looked upon to develop strategic ways of addressing/solving the problems, thereby contribute with suggestions to the successful business and economy or effective policy implementation.

The Institute promotes such projects through the grant of funds and provision of needful research infrastructure. Generally, the applied research projects are completed in the duration of six to eight months and they could be in the form of case study, monograph, organisation/firm-based study, evaluation-based study of policy/scheme/institution or other types of studies/projects as deemed necessary by the Faculty, provided they fit into the broad nature of applied research.

It is heartening to note that the initiative has been effectively grabbed by the Faculty who are recruiting students as research assistants in the process of the projects. It has been found that this exercise enriches the knowledge of the students by extending their academic activities, outside the classroom learning situation, in the real world.

The project outcome is intended to help the firm concerned in fixing up the problem/ addressing the given situation, and the Faculty to gain first-hand experience that enables in formulating hypotheses to get into a deeper research with wider scope. The findings from such practical exercises are disseminated to the wider world through FDPs, MDPs and publications. True to its objectives Faculty from SDMIMD are successful in harnessing the greater benefits of knowledge creation and its transfer from the applied research projects.

Dr.B.Venkatraja
Chairperson, SDM RCMS

Acknowledgement

At the outset, I would like to extend my earnest gratitude to SDME Trust for their constant encouragement and support in all my academic endeavours. I would like to sincerely thank, Dr. N. R. Parasuraman, Director, SDMIMD Mysore, for giving the opportunity to undertake this project. He has been a source of inspiration, and this work could not have been possible but for the encouragement and support of Dr. N. R. Parasuraman. I would also like to thank all the faculty and staff members, who have helped me directly or indirectly in completing this project.

Mohamed Minhaj

Table of Contents

Abstract	1
Introduction	1
Wikipedia: The World Knowledge Repository	3
Conceptual background of Text Analytics	12
Using Wikipedia's Knowledge	16
Wikipedia as a resource for Text Analytics	20
Conclusions and Future Work	30
References	31

1 Abstract

Text analytics has gained a great deal of attention in the recent years due to the tremendous amount of text data that is generated each day in a variety of forms and the business value that the organizations have identified in such textual sources(Chen et al., 2012). Among the plethora of textural sources available on the web, Wikipedia is one of the prominent text collections. Wikipedia has been the most successful collaborative encyclopaedia and is among the widely used websites around the globe. Wikipedia's English edition alone has around 6 million articles. However, the users are not able to leverage on the full potential of the vast knowledge base because of Wikipedia's constraints like limited full text search and machine interpretation. Hence, there are several research opportunities to study the use of Wikipedia's knowledge base directly or to use Wikipedia to support text analytics activities.

Today, most organisations are in the state of "Data rich and information poor". While a lot of data, predominantly unstructured and textual in nature is available in most organisations, they are not able to harvest and harness actionable information from that data. Text analytics deals with deriving novel, relevant and interesting patterns from data stored in organisations in the form of reports, web logs, emails, social media etc. Typical text analytics tasks include categorization, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization etc. Most of these tasks require access to large collection of updated and quality information either for handling the ambiguity associated with the natural language text or for the purpose of training and testing the Machine Learning models related to text analysis. In this context, Wikipedia has a great potential to be used for natural language processing and other applications of text analytics.

This project is an endeavour to study the content, structure and technology related to Wikipedia and to explore the possible ways of leveraging the vast, updated and collaboratively created knowledge of

Wikipedia for different text analytics tasks.

2 Introduction

The World Wide Web(WWW) is teeming with multiple forms of digital data. IDC, in a report sponsored by Seagate, predicts that the Global Datasphere will grow from 33 Zettabytes in 2018 to 175 Zettabytes by 2025(Reinsel et al., 2018). A significant amount of web-data is textual in nature and is available in the form of reports, news archives, scientific articles, blogs, emails, social networks etc. The WWW in its present form is a trove of knowledge, as it is being augmented by millions of people across the world. However, users are not able to effectively harvest and harness the vast amount of knowledge that is concealed in the web because of the limitations posed by the predominant genre of data being unstructured. Further, the sheer volume and heterogeneity makes the processing of text data a non-trivial task and requires efficient and effective techniques /algorithms to process the text on computers. While the unstructured text can be easily perceived by humans, it is significantly harder for machines to understand.

The increasing challenge of harvesting information from the vast amount of textual data on the web has fuelled the growth of research and development in Text Analytics. Text analytics is the process of deriving high-quality information from text. Text analytics generally involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text analytics refers to some combination of relevance, novelty, and interestingness. Typical text analytics tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization etc.

Among the plethora of textural sources available on the web, Wikipedia is one of the prominent sources. Wikipedia has been the most successful collaborative encyclopaedia and is among the widely used websites around the globe. Wikipedia's English edition alone

has around 6 million articles. Wikipedia users are not able to leverage on the full potential of Wikipedia's vast knowledge base because of Wikipedia's constraints like limited full text search and machine interpretation. Hence, there are several research opportunities to study use of Wikipedia's knowledge base directly or to use Wikipedia to support text analytics activities.

2.1 Objectives of the Proposed Research

1. Study the collaborative content authoring process, content structure and the underlying technology related to Wikipedia.
2. Study the various approaches to extract information from Wikipedia.
3. Study a sample of existing tools and applications, which have used Wikipedia to enhance their knowledge abilities.
4. Study the possible ways of leveraging the Wikipedia's knowledge base for text analytics.

2.2 Methodology

The focus of the current work is on the developments that have taken place in the research of Text Analytics in general and the use of the collaborative knowledge base of Wikipedia for Text Analytics in particular. The research endeavours to study the collaborative knowledge creation process of Wikipedia, knowledge structure and its extraction using exploratory method. Further, a Systematic Literature Review method has been employed for selecting, reviewing and synthesizing recent literature on possible ways of leveraging the Wikipedia's knowledge base for text analytics.

Systematic Literature Review (SLR) is a structured way of conducting a review of existing research works produced by the earlier researchers(Haneem et al., 2017). A systematic literature review attempts 'to identify, appraise and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a given research question (*About Cochrane*

Reviews / Cochrane Library, n.d.)

The systematic literature review is a method/process/protocol in which a body of literature is aggregated, reviewed and assessed while utilizing pre-specified and standardized techniques. In other words, to reduce bias, the rationale, the hypothesis, and the methods of data collection are prepared before the review and are used as a guide for performing the process. Just like it is for the traditional literature reviews, the goal is to identify, critically appraise, and summarize the existing evidence concerning a clearly defined problem(Štrukelj, 2018)

The Systematic literature reviews include all (or most) of the following characteristics:

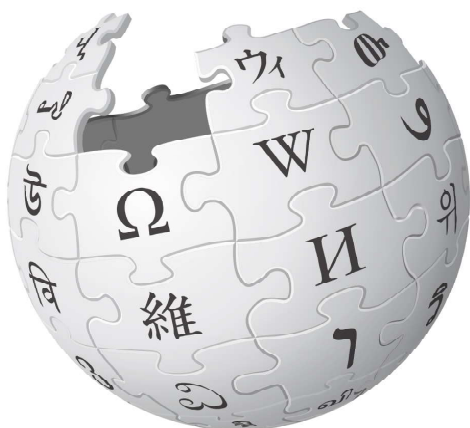
1. Objectives clearly defined a priori.
2. Explicit pre-defined criteria for inclusion/exclusion of the literature.
3. Predetermined search strategy in the collection of the information and systematic following of the process.
4. Predefined characteristic criteria applied to all the sources utilized and clearly presented in the review.
5. Systematic evaluation of the quality of the studies included in the review.
6. Identification of the excluded sources of literature and justification for excluding them.
7. Analysis/synthesis of the information (i.e., comparison of the results, qualitative synthesis of the results, meta-analysis).
8. References to the incoherences and the errors found in the selected material.

For the purpose of the review, the required scholarly literature pertaining to research objective was ferreted from Google Scholar. Google Scholar, a scholarly search engine widely used by the research community is a web search engine that is free and simultaneously indexes full text scholarly literature across many

disciplines and databases. Google Scholar indexes individual academic papers from journal and conference papers, theses and dissertations, academic books, pre prints, abstracts, technical reports and other scholarly literature from all broad areas of research (*The Use of Google Scholar for Research and Research Dissemination*, n.d.). While Google Scholar was the primary source used for searching the required literature, for the full text of the required articles that were not available in Google Scholar, appropriate sources like ACM, IEEE etc., were used.

The search terms to be used in Google Scholar were finalized based on an experimental basis. The data was filtered using the "Year of Publication". As the objective of the research is to study the recent developments related to the use of Wikipedia for Text Analytics, the literature extracted for analysis was of the period 2009 – 2019.

3 Wikipedia: The World Knowledge Repository



WIKIPEDIA

The Free Encyclopedia

Figure 1 Wikipedia - Logo

Wikipedia is a multilingual, web-based, free-content encyclopaedia project supported by

the Wikimedia Foundation and based on a model of openly editable content. The name "Wikipedia" is a portmanteau of the words wiki (a technology for creating collaborative websites, from the Hawaiian word wiki, meaning "quick") and encyclopaedia. Wikipedia's articles provide links designed to guide the user to related pages with additional information ("Wikipedia," 2019a).

Wikipedia is written collaboratively by largely anonymous volunteers who write without pay. Anyone with Internet access can write and make changes to Wikipedia articles, except in limited cases where editing is restricted to prevent disruption or vandalism. Users can contribute anonymously, under a pseudonym, or, if they choose to, with their real identity. The Wikipedia community has developed many policies and guidelines to improve the encyclopaedia; however, it is not a formal requirement to be familiar with them before contributing. The fundamental principles by which Wikipedia operates are the five pillars (Figure 2).

Wikipedia, the big data of text, facts and figures is open for all its users to edit and contribute to the fast-growing online repository. Wikipedia was founded as an offshoot of Nupedia, a now-abandoned project to produce a free encyclopaedia, begun by the online media company Bomis. Nupedia had an elaborate system of peer review and required highly qualified contributors, but the writing of articles was slow. During 2000, Jimmy Wales (founder of Nupedia and co-founder of Bomis), and Larry Sanger, whom Wales had employed to work on the encyclopaedia project, discussed ways of supplementing Nupedia with a more open, complementary project. Multiple sources suggested that a wiki might allow members of the public to contribute material, and Nupedia's first wiki went online on January 10, 2001. There was considerable resistance on the part of Nupedia's editors and reviewers to the idea of associating Nupedia with a website in the wiki format, so the new project was given the name "Wikipedia" and launched on its own domain, wikipedia.com, on January 15.

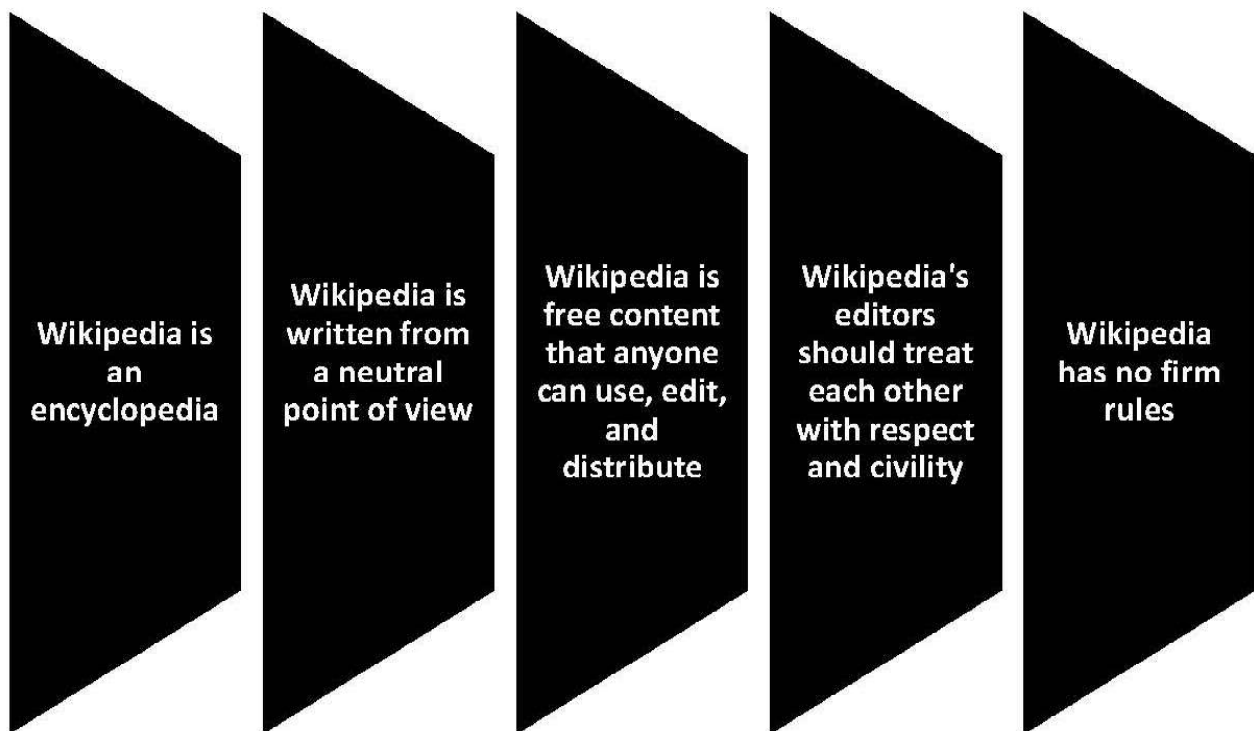


Figure 2 Five Pillars of Wikipedia

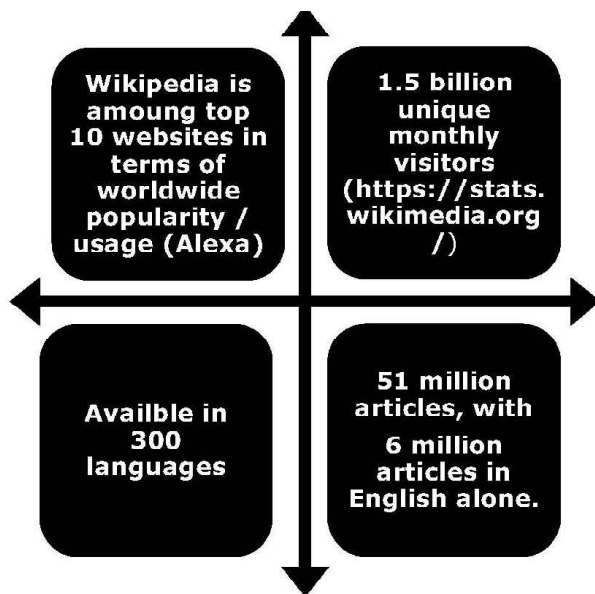


Figure 3 Interesting facts about Wikipedia

directly from a web browser. In a typical wiki, text is written using a simplified mark-up language and often edited with the help of a rich-text editor("Wiki," 2020)

Wikis run using wiki software, known as wiki engine. A wiki engine may be defined as a type of a CMS (Content management System), but there lies a lot of differences compared to other CMS, because of the elements like "blog software", this means the content creation part is not bound by any owner and wikis have little implicit structure, allowing structure to emerge according to the needs of the users. There are several wiki engines in use today, either as a standalone or part of some other software. Some wiki engines are proprietary in nature, whereas others are open source.

The online encyclopaedia project Wikipedia is by far the most popular wiki-based website and is one of the most widely viewed sites of any kind in the world and has been ranked in the top ten since 2007.

3.1 What is Wiki ?

A wiki is a knowledge base website on which users collaboratively modify and structure content

3.2 The Technology behind Wikipedia:

The major operations of Wikipedia are facilitated by MediaWiki, which is a custom-made, free and open source wiki software platform written in PHP. MediaWiki is the technology backbone of many prominent websites including, Wiktionary and Wikimedia Commons. While the database backend used for the systems is MySQL and MariaDB, Nginx and Varnish are used for the frontend and caching. The Wikipedia system is deployed as a web application using Apache as the Web Server. Wikipedia and the other Wikimedia projects are configured on several racks full of servers running primarily on Linux, Debian distribution ("Wikipedia," 2020).

In addition to employing the above technology that is facilitating its basic operations, Wikipedia is at the technology forefront when it comes to providing effective and efficient services to its users. Some of the notable technologies used for different aspects of improving the quality of its services are as follows:

3.2.1 Objective Revision Evaluation Service (ORES)

ORES is an AI based service developed by Wikimedia Foundation to identify mischievous edits to Wikipedia articles and address them quickly. ORES uses machine learning strategies to "learn" what damaging edits look like and flag them for review by automated tools or human editors. ORES has helped cut the workload of Wikipedia volunteers by a factor of 10—reducing the effort of combating vandalism from 270 human hours per day to a mere 27 ("The technology behind free knowledge," 2019).

3.2.2 Page preview popup

One of the key features of Wikipedia is Wikification. Wikification is the process of providing background knowledge about the topic the user is reading by enabling links to other Wikipedia pages pertaining to the terms that the user encounters while reading about a particular topic. However, this process involves navigating away from the original Wikipedia page and

returning to the page after reading the secondary or background information. Considering these usability and readability issues, a new feature was built by Wikipedia developers to give the users a quick grasp of what is behind a link without leaving the topic that the user is reading. This feature enables the ability to learn more with fewer distractions.

3.2.3 Content translation tool

Wikipedia has long aspired to foster a world where knowledge is free, accessible and useful to everyone, in every language, in every country, and across any device ("The technology behind free knowledge," 2019). Towards this end, translation across Wikipedias in hundreds of languages is critically important: it allows multilingual editors to re-use efforts made by other volunteer editors, thereby lowering the cost of spreading knowledge around the world. Considering this requirement, a content translation tool was built, which simplified the translation process and facilitated in bringing the sum of all knowledge into other languages. A part group of volunteer translators focused on making vital medical information available in multiple languages reported that the translation tool increased their productivity by 17 % ("Wikipedia's coverage of essential vaccines is expanding," 2016).

3.3 Wikipedia Content

Wikipedia, a multilingual, web-based, free-content encyclopaedia is based on a model of openly editable content. Since its creation in 2001, Wikipedia has grown rapidly into the world's largest reference website, attracting 1.5 billion unique visitors monthly (as of March 2020). Unlike printed encyclopaedias, Wikipedia is continually created and updated, with articles on historic events appearing within minutes, rather than months or years. Wikipedia has become more comprehensive than any other encyclopaedia because everybody can help improve it. The Wikipedia contributors in addition to the quantity, work on improving the quality as well. Every article on Wikipedia is considered to be in a work-in-progress phase and progresses to various stages of completion. As articles

develop, they tend to become more comprehensive and balanced. Quality also improves over time as misinformation and other errors are removed or repaired. However, because anyone can click “edit” at any time and add content, any article may contain undetected misinformation, errors, or vandalism and hence due diligence is crucial while using the content.

3.3.1 Content License:

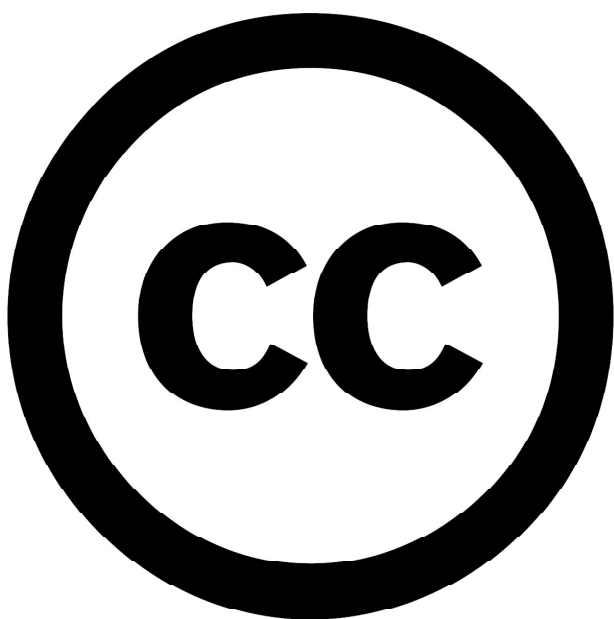


Figure 4 Creative Commons License - Logo

The Wikipedia content is available under the Creative Commons Attribution-ShareAlike License. A Creative Commons (CC) license is one of several public copyright licenses that enable the free distribution of an otherwise copyrighted work (*The teacher’s guide to Creative Commons licenses / Open Education Europa*, n.d.) A CC license is used when an author wants to give other people the right to share, use, and build upon a work that they have created. CC provides an author flexibility (for example, they might choose to allow only non-commercial uses of a given work) and protects the people who use or redistribute an author’s work from concerns of copyright infringement as long as they abide by the conditions that are specified in the license by which the author distributes the work.

3.3.2 Content Classification:

Wikipedia’s content is classified into several categories. Each category has several subcategories. When a user writes an article, his article is classified under these categories and their sub-categories. Users are expected to mention the category while creating the page using the keyword “Category”. A page can be added to more than one category. The categories are displayed in the bottom of the page. The categories aid the users in browsing content related to a particular domain or search content pertaining to a specific category or subcategory.

Table 1 Wikipedia Content Categories

Reference works – compendiums of information, usually of a specific type, compiled in a book for ease of reference. That is, the information is intended to be quickly found when needed.
Culture – encompasses the social behaviour and norms found in human societies, as well as the knowledge, beliefs, arts, laws, customs, capabilities and habits of the individuals in these groups.
Geography – field of science devoted to the study of the lands, features, inhabitants, and phenomena of the Earth and planets.
Health – state of physical, mental and social well-being in which disease and infirmity are absent.
History – the past as it is described in written documents, and the study thereof.
Human activities – the various activities done by people. For instance, it includes leisure, entertainment, industry, recreation, war, and exercise.

Mathematics – the study of topics such as quantity (numbers), structure, space, and change. It evolved using abstraction and logical reasoning, from counting, calculation, measurement, and the systematic study of the shapes and motions of physical objects.
Natural science – branch of science concerned with the description, prediction, and understanding of natural phenomena, based on empirical evidence from observation and experimentation.
People – plurality of persons considered as a whole, as is the case with an ethnic group or nation.
Philosophy – study of general and fundamental questions about existence, knowledge, values, reason, mind, and language.
Religions – social-cultural systems of designated behaviours and practices, morals, worldviews, texts, sanctified places, prophecies, ethics, or organizations, that relates humanity to supernatural, transcendental, or spiritual elements.
Society – group of individuals involved in persistent social interaction, or a large social group sharing the same geographical or social territory, typically subject to the same political authority and dominant cultural expectations. Societies are characterized by patterns of relationships (social relations) between individuals who share a distinctive culture and institutions; a given society may be described as the sum total of such relationships among its constituent of members.
Technology – the sum of techniques, skills, methods, and processes used in the production of goods or services or in the accomplishment of objectives, such as scientific investigation.

For example, the wiki page of Dr. Veerendra Heggade has been tagged with several categories like, Living People, Indian Philanthropist, recipient of Padma Vibhushan etc.



The screenshot displays the Wikipedia page for Veerendra Heggade. The page includes a sidebar with navigation links, a main content area with the article text, and a list of categories at the bottom. The categories listed are: Living people, Indian philanthropists, Recipients of the Padma Bhushan in social work, 20th-century Indian Jains, People from Dakshina Kannada district, Mangaloreans, Tulu people, 1948 births, Recipients of the Karnataka Ratna, Recipients of the Padma Vibhushan in social work, Heggade family, 21st-century Indian Jains, Indian male social workers, and Social workers from Karnataka.

Figure 5 Example of Page Categories : Wiki page of Dr. Veerendra Heggade.

3.3.3 Infoboxes: Structured data enveloped in the textual content of Wikipedia articles.

An Infobox is a fixed-format table usually added to the top right-hand corner of articles to consistently present a summary of some unifying aspect that the articles share and sometimes to improve navigation to other interrelated articles. The use of infoboxes is neither required nor prohibited for any article. Whether to include an infobox, which infobox to include, and which parts of the infobox to use, is determined through discussion and consensus among the editors at each individual article.

Infoboxes contain important facts and statistics of a type which are common to related articles. They are like fact sheets, or sidebars, in magazine articles. They quickly summarize important points in an easy-to-read format. However, they are not "statistics" tables and only summarize material from an article—the information should still be present in the main text, partly because it may not be possible for some readers to access the contents of the infobox. If infobox templates hide long columns of data inside collapsing tables, then readers using assistive technology may miss the content.

Considering the structured nature of data and possibility of mapping its schema easily to many prominent metadata systems, the data in Wikipedia's infoboxes is widely used for many knowledge-based applications. The notable applications that are primarily built on Wikipedia's infoboxes include DBpedia.

Infosys Limited	
Infosys	
Type	Public
Traded as	BSE: 500209 ^g NSE: INFY ^g NYSE: INFY ^g BSE SENSEX Constituent CNX Nifty Constituent
ISIN	US4567881085 [✓]
Industry	IT services, IT consulting
Founded	7 July 1981; 38 years ago
Founders	N.R. Narayana Murthy Nandan Nilekani S. Gopalakrishnan S. D. Shibulal K. Dinesh N. S. Raghavan Ashok Arora
Headquarters	Bangalore, Karnataka, India
Area served	Worldwide
Key people	Nandan Nilekani (Chairman) Salil S. Parekh (MD & CEO) ⁽¹⁾
Services	Outsourcing · Consulting · Managed services
Revenue	▲ US\$12.4 billion (2019) ⁽²⁾
Operating income	▲ US\$2.7 billion (2019) ⁽²⁾
Net income	▼ US\$2.2 billion (2019) ⁽²⁾
Total assets	▼ US\$12.2 billion (2019) ⁽²⁾
Total equity	▼ US\$9.4 billion (2019) ⁽²⁾
Number of employees	243,454 (Dec 2019) ⁽²⁾
Divisions	Infosys BPM EdgeVerve Systems Infosys Consulting
Website	www.infosys.com ^g
Footnotes / references	
⁽²⁾	

Figure 6 Infobox on Wikipedia related to InfoSys

3.4 Creating articles in Wikipedia

Besides accessing the required information from one of the largest knowledge bases available on the web, the Wikipedia users can enhance the knowledge base by contributing additional information. This collaborative approach benefits the entire Wikipedia community.

3.4.1 What kind of articles can the user contribute?

- A user may write about a Place, Person, Organization, Technology or could be about anything and everything in the world.
- Only concern is the article should be referred/ cited from authenticated sources.
- Sources which exercise some form of editorial control are best suited to be used as references such as: newspapers, books, journals, magazines, etc.

The goal of a Wikipedia article is to create a comprehensive and neutrally written summary of existing mainstream knowledge about a topic. Articles should have an encyclopaedic style with a formal tone instead of essay-like, argumentative, promotional, or opinionated writing.

3.4.2 The key points that the contributor should keep in mind are:

- ◆ Decision to create a page about any topic should be taken only if it is about a notable topic and it does not violate copyright law.
- ◆ Wikipedia is an encyclopaedia, and not a personal home page or a business list.
- ◆ For the topic to be considered notable, it must be covered in detail in good references from independent sources.

- ◆ Copy-paste of content from other websites even if it is owned by the user, is not encouraged.
- ◆ To create an article, the user account must be at least 4 days (96 hours) old and is expected to have made more than 10 edits.
- ◆ If the topic is not notable, or contains copyrighted material, the article will be rejected or deleted by the Wikipedia Community.

3.4.3 The steps involved in creating a page in Wikipedia:

- ◆ The first step in creating a page involves searching the topic in Wikipedia itself to check whether an article with the same title exists or not.
- ◆ If the article already exists, the user can edit the article if he possesses some additional information about the topic.
- ◆ If the article doesn't exist, Wikipedia provides two interfaces for creating a page:
 - Source Editing – Facilitates creation of pages using Wikitext. Wikitext, also known as Wiki markup or Wikicode, consists of the syntax and keywords used by the MediaWiki software to format a page.
 - Visual Editing – Facilitates creation of pages using *What You See Is What You Get* editor. This Visual Editor was created by the Wikimedia Foundation in collaboration with Wikia. The visual editing facility was enabled to reduce the technical difficulty in compiling the pages and thereby encouraging more users to contribute.

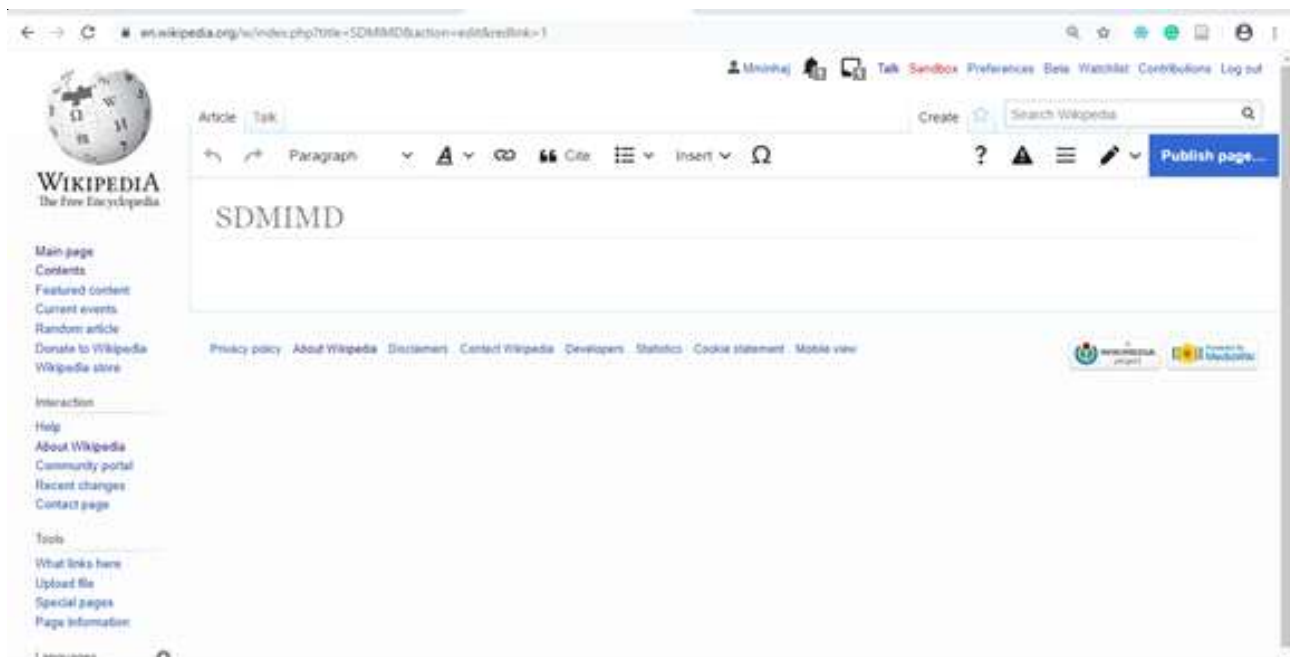


Figure 7 Wikipedia - Page Editor

Wikipedia provides several options to hand hold the novice contributors:

- First, a sand box system is provided for the contributors to practice the Wikipedia editing without affecting live articles.
- Page Creation Wizard is provided for the contributors to carefully follow the step by step procedure to create a page. Adhering to the guidelines mentioned in each of these steps ensure that the possibility of the page getting deleted after the review is reduced.
- Further, Special:Mypage option is available to reduce the risk of deletion as other editors help in developing the article. The article is moved to the "Article Space" only after it is ready.
- Every page on Wikipedia has an associated Talk page. Talk pages also known as discussion pages are administrative pages where editors can discuss improvements to articles or other Wikipedia pages. The talk page associated with an article is named "Talk:Example", where "Example" is the name of the article. For example,

the talk page for discussion of improvements to the article India is named Talk:India.

- If a Wikipedia user has a question, concern, or comment related to improving a Wikipedia article, he can put a note on that article's talk page.

3.4.4 How to Insert Media Files?

Images, sounds, and videos enhance articles greatly. Wikipedia supports several types of file format including png, gif, jpg/jpeg, xcf, pdf, mid, ogg/ogv/oga, svg, djvu with maximum size of 100 megabytes. The files must be basically uploaded to Wikipedia Commons. This allows the files to be used in articles instantly by all Wikimedia projects. A file that is already hosted on the Wikimedia Commons can be inserted into a Wikipedia Page with the code `[[File:FILENAME|thumb|DESCRIPTION]]`.

3.4.5 Wikipedia Protection Policy

Wikipedia is built around the principle that anyone can edit it and therefore most of its pages are open for public editing. However, certain pages are restricted for editing. In some circumstances, because of a



Figure 8 Upload Media Page in Wikipedia

specifically identified likelihood of damage resulting if editing is left open, Wikipedia restricts the editing of certain pages. These restrictions in the form of “Protection” are handled by Wikipedia administrators only, therefore the removal or application lies with them. These restrictions can be time bound or for an infinite period. “Full-Protection” is the most commonly used protection, which means that the page can only be edited by the administrators, whereas “Semi-protection” means that a page could also be modified by logged-in users. There are many other variants of the content protection.

3.5 Content Assessment

The Wikipedia’s content assessment is performed to check the quality of articles and make sure that the standard is maintained for all the articles. The articles are categorized according to the official review and the quality of content, stages of completion, etc.

3.6 The Review Process

Whenever a Wikipedia user creates a new article, a volunteer editor reviews the article and moves it to the main space or provide comments to the creating editor on how to improve the article. If an article is “approved,” it is moved to the main space as a live article and becomes subject to all the guidelines applicable to every other article in Wikipedia. That means that it can be edited by anyone in the

	Fully protected
	Template-protected
	Semi-protected
	Create protected
	Move protected
	Upload protected
	Pending changes protected
	Extended confirmed protection
	Protected by Office
	Cascade protected

Figure 9 Different types of Wikipedia Content Protection

community and also become subject to deletion review if an editor feels the article is not notable. So, even if the article is initially “approved”, it can still be deleted by the community.

Wikipedia employs Peer - Review process for getting feedback about the articles. The process is very open, any editor can give feedback and the contributor can

Given below is the official list by Wikipedia for the article categorisation:

Class	Criteria	Reader's experience	Editing suggestions	Example
FA	The article has attained featured article status. More detailed criteria [show]	Professional, outstanding, and thorough; a definitive source for encyclopedic information.	No further content additions should be necessary unless new information becomes available; further improvements to the prose quality are often possible.	Fluorine (as of June 2017)
A	The article is well-organized and essentially complete, having been reviewed by impartial reviewers from a WikiProject, like military history , or elsewhere. Good article status is not a requirement for A-Class. More detailed criteria [show]	Very useful to readers. A fairly complete treatment of the subject. A non-expert in the subject matter would typically find nothing wanting.	Expert knowledge may be needed to tweak the article, and style issues may need addressing. Peer review may help.	Fluorine (as of March 2012)
GA	The article has attained good article status. More detailed criteria [show]	Useful to nearly all readers, with no obvious problems; approaching (although not equalling) the quality of a professional encyclopedia.	Some editing by subject and style experts is helpful; comparison with an existing featured article on a similar topic may highlight areas where content is weak or missing.	Osmium (as of March 2012)
Bplus	Detailed, clear and accessible, often with history or images; possible good article nominee . More detailed criteria [show]	Useful to nearly all readers. A good treatment of the subject which attempts to be as accessible as possible, with a minimum of jargon. No obvious problems, gaps, excessive information.	May be improved by input from experts to assess where coverage is still missing, and also by illustrations, historical background and further references. Consider peer review or nominating for good article status . If the article is not already fully wikified , now is the time.	Antimony (as of March 2012)
B	The article is mostly complete and without major issues, but requires some further work to reach good article standards . More detailed criteria [show]	Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.	A few aspects of content and style need to be addressed. Expert knowledge may be needed. The inclusion of supporting materials should also be considered if practical, and the article checked for general compliance with the Manual of Style and related style guidelines .	Gold (as of March 2012)

request feedback that is more specific than generic. The important points to be kept in mind by the contributor of the page regarding review process are as follows :

- § For a submitted article to be reviewed, 14 days must have passed since the previous peer review of that article.
- § For nominating a page for review, addition of {{subst:PR}} to the top of the article's talk page is required.
- § The peer review will be automatically listed within an hour, when page ends with four tildes (~ ~ ~ ~).
- § If article contributor has received minimal feedback, or if the review is edited more than once, the page can be manually added to the backlog list. This ensures reviewers don't overlook the review request.

4 Conceptual background of Text Analytics

In the present milieu of business, organisations are generating data at incredible speed and volume. The digitization of business processes is enabling the organisations in capturing finer details of all aspects of their business, like their employees, customers and their behaviours. The organisations are constantly looking for ways to use the captured data to address their business problems and differentiate themselves in the market. But with the paradigm shift in the data growth from mostly structured to mostly unstructured data, organisations are facing challenges in leveraging the stored data. It is estimated that 80% of the world's data is unstructured(*The biggest data challenges that you might not even know you have*—Watson, n.d.), but businesses are only able to gain visibility into a portion of that data. Businesses use structured data every day through relational databases and

spreadsheets, where patterns can easily be identified. However, unstructured data, which comes in the form of emails, social media, blogs and documents is trove of business opportunities. Due to its unstructured nature, it is difficult for organisations to gain insights from it using conventional systems. And because so much of data created today is unstructured, organizations need to be able to understand what is in this data, or risk missing out on significant amounts of digital intelligence in the form of opinions, customer sentiments and feedback. This phenomenon has increased the significance of text analytics in organisations. This section of the report endeavours to present the key concepts related to Text Analytics in the light of scholarly literature. The text presented in this section of the report is intended to serve as a broad, conceptual background about Text Analytics for exploring different techniques and use cases of leveraging Wikipedia for text analytics.

With the advent of huge amount of textual data that the organisations have and the value that organisations have realised in harvesting and harnessing their textual data, Text Analytics has gained a lot of prominence in the recent times. However, text analytics poses some challenges compared to regular analytical methods. Free flowing text is highly unstructured and rarely follows any specific pattern. This limitation increases the complexity of employing the standard analytical techniques and statistical methods for analysing the textual data. Text analytics demands a different approach compared to data analytics and has evolved as a new discipline of research and development.

Text analytics also known as text mining or knowledge discovery from text is the methodology and process followed to derive quality and actionable information and insights from textual data. Text mining was first introduced by Fledman et al. (Feldman & Dagan, 1995). It involves using Natural Language Processing (NLP), Information Retrieval (IR) and Machine Learning (ML) techniques to parse unstructured text data into more structured forms and deriving patterns and insights from this data that would be helpful for the end user. Text analytics comprises a collection of machine

learning, linguistic and statistical techniques that are used to model and extract information from text primarily for analysis needs, including business intelligence, exploratory, descriptive and predictive analytics (Sarkar, 2016).

4.1 Text representation and encoding

Text mining from a large textual documents is a complex process, thus it is critical to have a data structure for the text which facilitates further analysis of the documents (Hotho et al., n.d.). The most common approach to represent the textual data is Bag of Words (BoW). A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things, a vocabulary of known words/terms and a measure of the presence of known words/terms. It is called a “bag” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document. This representation leads to a vector representation which is referred as Term by Document Matrix (TDM) and can be used to analyze the text using machine learning and statistics.

	Term 1	Term2	Term 3	...	Term M
Document 1					
Document 2					
Document 3					
...					
Document N					

Table 2 Term by Document Matrix

4.2 Dimension Reduction Techniques

One major problem encountered with BoW approach is high dimensional data. With all the words/terms present in all the documents being added as columns (features) to the TDM, the dimensionality of the table would grow exponentially. To address this issue, besides stemming and elimination of stop words, many dimension reduction techniques are used. The

prominent techniques used for dimension reduction include Principal Components Analysis (PCA), Metric Multidimensional Scaling (MDS) etc.

PCA is a correlation-based technique that chooses a set of representative dimensions called the principal components based on the degree of variation that they capture from the original data. In particular, PCA computes the output points by performing a singular value decomposition (SVD) on the document covariance matrix and then multiplies the resulting eigenvectors with their corresponding eigenvalues (Underhill et al., 2007).

MDS focuses on preserving distances between pairs of points (Wickelmaier, 2003). The input is a matrix containing pairwise distances between the original points. MDS performs an eigenvalue decomposition on this matrix in a way that embeds the points in a smaller space while maintaining the relative pairwise relationships.

4.3 Text Preprocessing

Preprocessing is one of the key tasks in Text Analytics and has high impact on the quality of any text analytics outcome. For example, a traditional text categorization framework comprises preprocessing, feature extraction, feature selection and classification steps. While it is evident from several studies that feature extraction, feature selection and classification algorithm have impact on the classification process, the preprocessing stage has exhibited significant influence on the success of text categorization. Uysal et al. have investigated the impact of preprocessing tasks particularly in the area of text classification (Uysal & Gunal, 2014). The preprocessing step usually consists of the tasks such as tokenization, filtering, lemmatization and stemming.

Tokenization: It is the first step in text analytics. The process of breaking down a text into smaller chunks such as words or sentence called tokens is called Tokenization. The list of tokens is then used for further processing of text.

Filtering: It refers to removal of certain words from

the text being processed. The most common filtering is stop-words removal. Stop words are the words that appear frequently but are very generic in meaning and do not have much content information (Ex. prepositions, conjunctions, etc.).

Stemming: It is the process of eliminating affixes (suffixed, prefixes, infixes, circumfixes) from a word in order to obtain a word stem. Ex. running -> run.

Lemmatization: Lemmatization is related to stemming and is focused on capturing canonical forms based on a word's lemma. It involves grouping together the various inflected forms of a word so they can be analyzed as a single item. In other words, lemmatization methods try to map verb forms to infinite tense and nouns to a single form.

4.4 Text Analytics Approaches:

Information Retrieval (IR): It is the activity of finding information resources (usually documents) from a collection of unstructured data sets that satisfies the information need (*A Survey of Information Retrieval and Filtering Methods*, n.d.). Information Retrieval deals with facilitation of information access rather than analyzing information and finding hidden patterns, which is the main purpose of mining.

Natural Language Processing (NLP): It is a field of study which combines computer science, artificial intelligence and linguistics with an aim of understanding the natural language using computers (Liddy, n.d.). Many of the text mining algorithms extensively make use of NLP techniques, such as parts of speech tagging (POS), syntactic parsing and other types of linguistic analysis.

Text Summarization: Many text mining applications need to summarize the text documents in order to get a concise overview of a large document or a collection of documents on a topic (Radev et al., 2002). There are two categories of summarization techniques in general: extractive summarization where a summary comprises information units extracted from the original text, and in contrary abstractive summarization where a summary may contain "synthesized" information that may not occur in the original

document (Allahyari et al., 2017).

Information Extraction from text (IE): Information Extraction is the task of automatically extracting information or facts from unstructured or semi-structured documents (Cowie & Lehnert, 1996). It usually serves as a starting point for other text mining algorithms (Allahyari et al., 2017). For example, extraction of entities and their relations from text can give us useful semantic information.

Machine Learning (ML): It is a subset of AI that provides computing systems with the ability to learn without being explicitly programmed. ML focuses on the development of Computer Programs that can change when exposed to new data. As far as analytics is concerned, ML is a method of data analysis that automates analytical model building and works on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Any type of machine learning requires past figures, facts and experiences given in the form of data for it to predict or act on the new set of data. The data that is used for this purpose of training the machine is called Training Data and primarily contains set of examples or background knowledge to facilitate learning.

Based on the different approaches used to train the machines or in other terms the way the system is modeled to discover the patterns from data, machine learning can be broadly classified as Supervised and Unsupervised.

Supervised learning: When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples, such type of machine learning is referred as Supervised ("An introduction to Machine Learning," 2017).

Ex. Text Classification

Unsupervised Learning: The type of ML where an algorithm learns from plain examples without any associated response, leaving to the algorithm to

determine the data patterns on its own, is referred as unsupervised learning.

Ex. Text Clustering, Topic Modeling

4.4.1 Text Classification :

Text classification also known as text tagging or text categorization is the process of categorizing text into organized groups. By using Natural Language Processing (NLP), text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content (*What is Text Classification?*, n.d.).

Common example of Text Classification include :

§ Sentiment Analysis: The process of understanding, if a given text is talking positively or negatively about a given subject (e.g. for brand monitoring purposes).

§ Topic Detection: The task of identifying the theme or topic of a piece of text (e.g. know if a product review is about Ease of Use, Customer Support, or Pricing when analyzing customer feedback).

§ Language Detection: The procedure of detecting the language of a given text (e.g. know if an incoming support ticket is written in English or Spanish for automatically routing tickets to the appropriate team).

4.4.2 Text Clustering:

It is the application of cluster analysis to text-based documents. It uses machine learning and natural language processing (NLP) to understand and categorize unstructured, textual data. Typically, descriptors (sets of words that describe topic matter) are extracted from the document first. Then they are analyzed for the frequency in which they are found in the document compared to other terms. After which, clusters of descriptors can be identified and then auto tagged.

4.4.3 Topic Modeling

This provides a simple approach to analyse large amount of data by identifying the hidden thematic structures in document collections. The Topic models help in annotating documents according to the inherent topics and aids in summarizing and searching. Topic models are currently being used for a variety of purposes like enhancing search facility, recommendation systems, classification of information assets etc.

5 Using Wikipedia's Knowledge

While the use of Wikipedia for scholarly activities is a subject of debate, it is being used extensively by all types of web users, from students to professors, interns to CXOs, as an easily accessible tertiary source for information about anything and everything, as a quick

"ready reference", to get a sense of a concept or idea("Wikipedia," 2019).

The concerns expressed by some users about the authenticity of the Wikipedia content is primarily because of its open and collaborative nature. The content on Wikipedia could contain vandalism and wrong information and hence may not be suitable for academic or formal research activities. However, Wikipedia practises a rigorous review process and has implemented a strong content monitoring system. Furthermore, Wikipedia is not a platform for creation of new knowledge and instead is a vast collection of knowledge gathered from numerous peer reviewed information sources and submitted by numerous users across the world. The following section presents the different ways of using the Wikipedia's knowledge base.

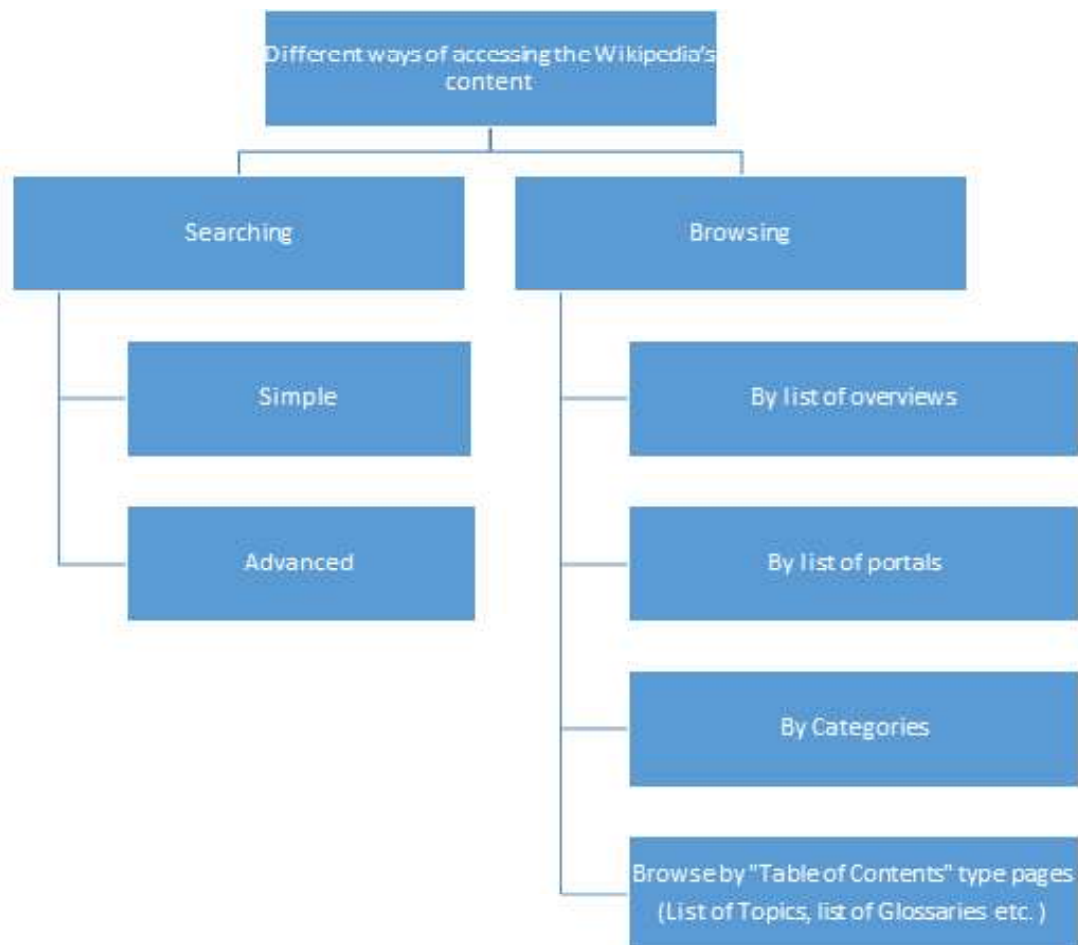


Figure 10 Different ways of accessing the Wikipedia's Content

5.1 The simple search option:



Figure 11 Homepage of Wikipedia (<https://www.wikipedia.org>)

The simple search option provided on the Wikipedia's homepage is one of the most widely used facility to access the Wikipedia's information. The text box provided for performing the search is a versatile option and provides facility to search in multiple languages, in addition to having auto – complete /

auto – suggest feature.

In addition to the above-mentioned search facility available on the home page, a search box is available on all the pages of Wikipedia. This option facilitates searching for a specific entity page or pages containing the search term in the full-text of the page.



Figure 12 Search Box on all pages of Wikipedia

5.2 Advanced Search:

Furthermore, an advanced search option is provided by Wikipedia to facilitate field specific search and filters to refine the search results. The figure 13, depicts the different options provided on advanced search page.

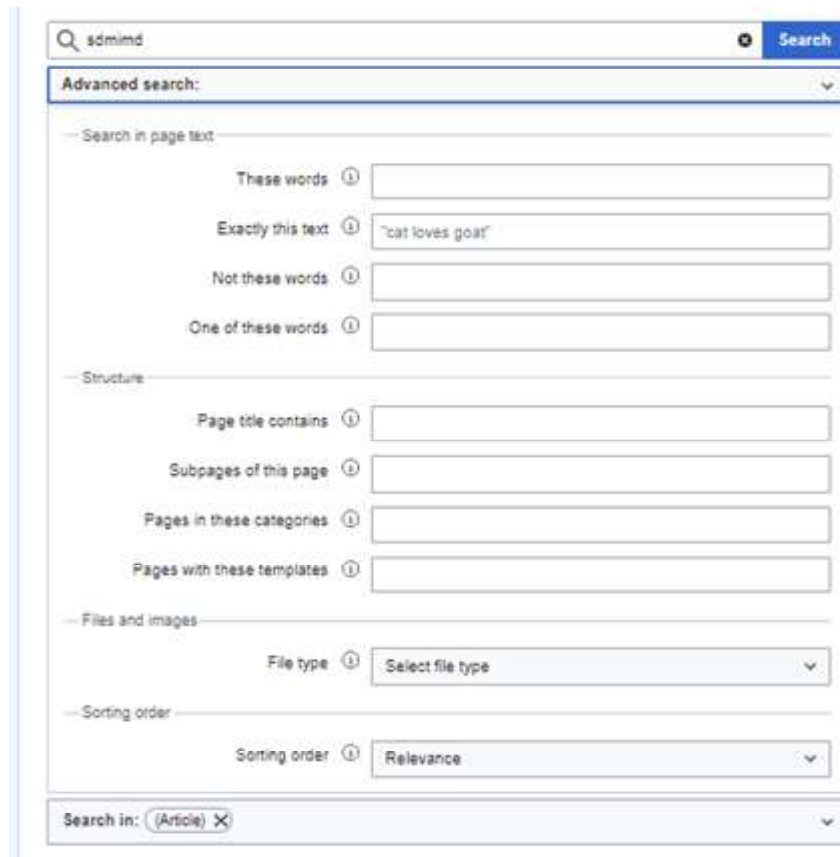


Figure 13 Wikipedia - Advanced Search

5.3 Overviews:

An overview is a survey of what is covered or included in a subject area of Wikipedia. The outline of Wikipedia

is divided into sections, each providing an overview of a major subject area, presenting key articles in the respective subject is given in the overviews ("Wikipedia," 2019).

5.4 Portals :

They complement main topics in Wikipedia and expound upon topics by introducing the reader to key articles, images, and categories that further describe the subject and its related topics. Portals also assist in helping editors to find related projects and things they can do to improve Wikipedia and provide a unique way to navigate Wikipedia topics. As on February 2020, there are 528 portals.

5.5 Categories :

They help the Wikipedia user to find information, even if he does not know what exists or what it is called. The list of categories of Wikipedia's coverage parallels their other lists by topic.

General reference
Culture and the arts
Geography and places
Health and fitness
History and events
Human activities
Mathematics and logic
Natural and physical sciences
People and self
Philosophy and thinking
Religion and belief systems
Society and social sciences
Technology and applied sciences

Figure 14 Wikipedia Content Overviews



Figure 15 Example of Wikipedia Content Category (Culture and the arts)

5.6 Retrieving information from Wikipedia using Search Engines:

In addition to the search facility available on Wikipedia to retrieve the required information, the popular search engines like Google can also be used as an interface to get information from Wikipedia. First, as depicted in figure 15, in most cases, Wikipedia is top

of the list in the Google's search engine results page and Google's knowledge graph rendered in the form of infobox also uses a significant amount of data from Wikipedia. Secondly, as depicted in figure 16, Wikipedia's information can be explicitly search from google by using a term "wiki" as prefix or suffix in the search query.

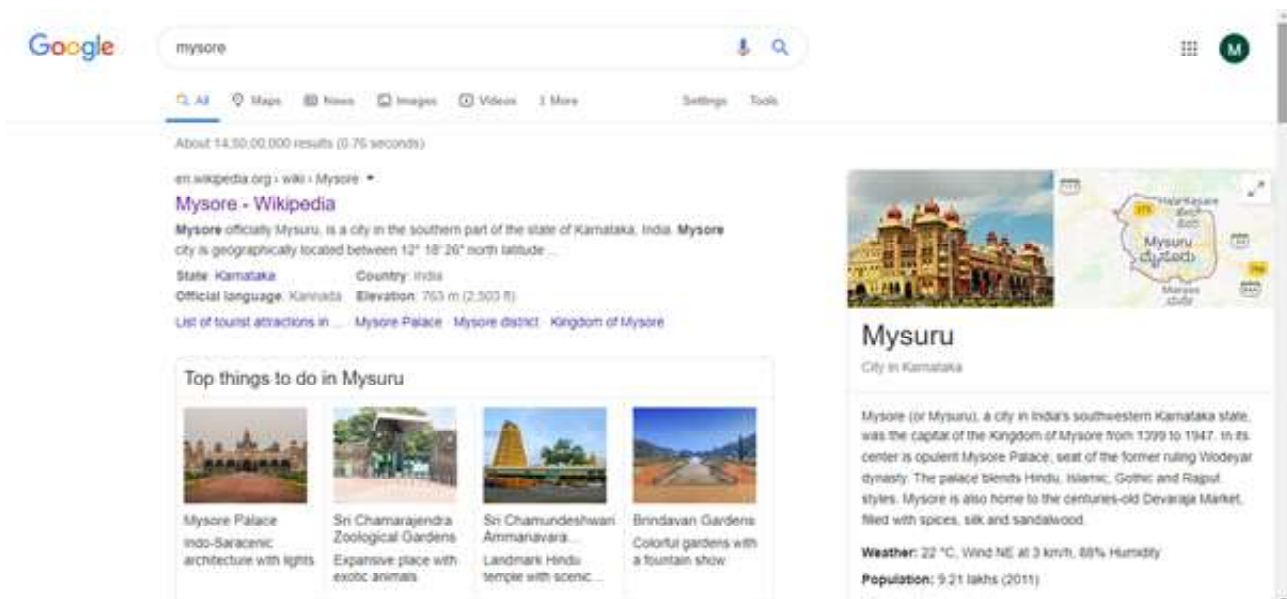


Figure 16 Google's results page with Wikipedia page on top of the list and Wikidata being used in the Google's infobox

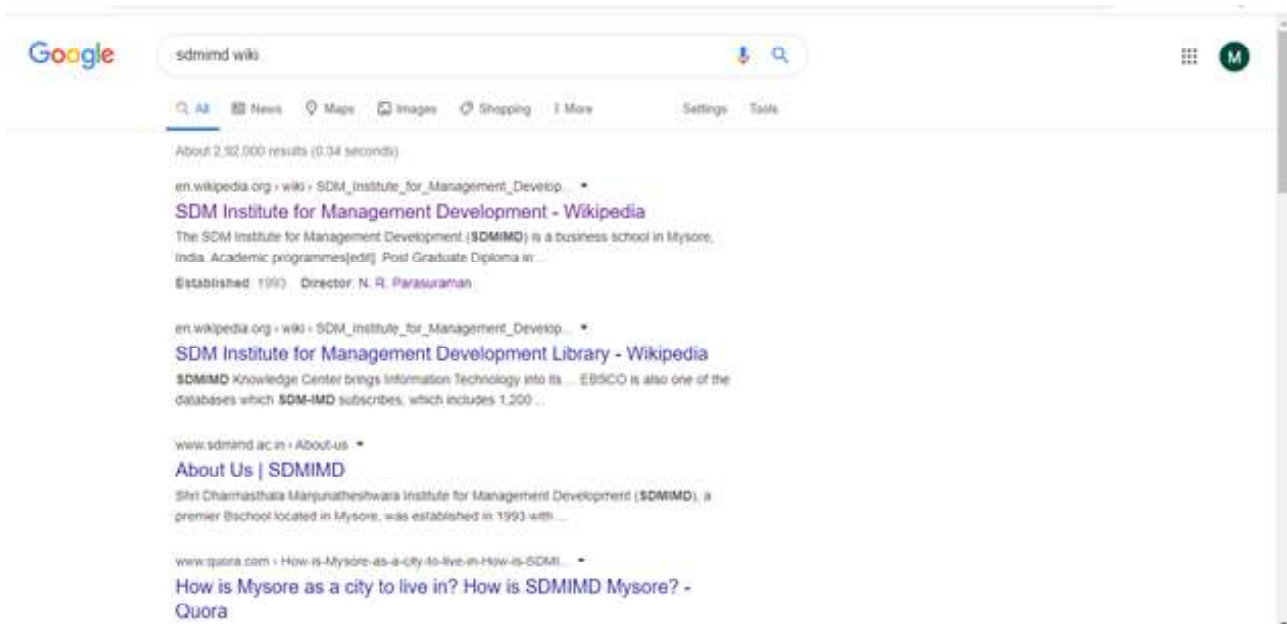


Figure 17 Use of wiki prefix/suffix in search query

6 Wikipedia as a resource for Text Analytics

Text Analytics refers to the discipline of Computer Science which employs Natural Language Processing and Machine Learning to draw meaning from unstructured text data. Numerous applications of Text

analytics include, business analyst turning thousands of guest reviews into specific recommendations, workforce analyst improving productivity and reducing employee turnover and healthcare provider understanding patient experiences (Mohler, 2019).

Getting business insights from textual data involves several steps such as tokenization, removal of stop words, lemmatization, POS tagging and use of statistical methods and machine learning. Text analytics tasks include Information Retrieval, Information Extraction, Text Summarization, Text Classification etc.

The data used by the organisations for the above-mentioned tasks include emails, web logs, reviews, feedback, survey data, data harvested from social networks such as Facebook, Twitter etc. However, many of the text analytics tasks require access to large amount of textual data as secondary or background

data either for the purpose of understanding the text being analysed or for training machine learning models. In such a context, vast amount of collaboratively curated knowledge of Wikipedia can play a pivotal role.

In the light of recent scholarly literature, the works involving Wikipedia for any form of Text Analytics have been explored and presented in the following section. The different categories of text analytics tasks that have leveraged on Wikipedia knowledge have been depicted in the figure 18.

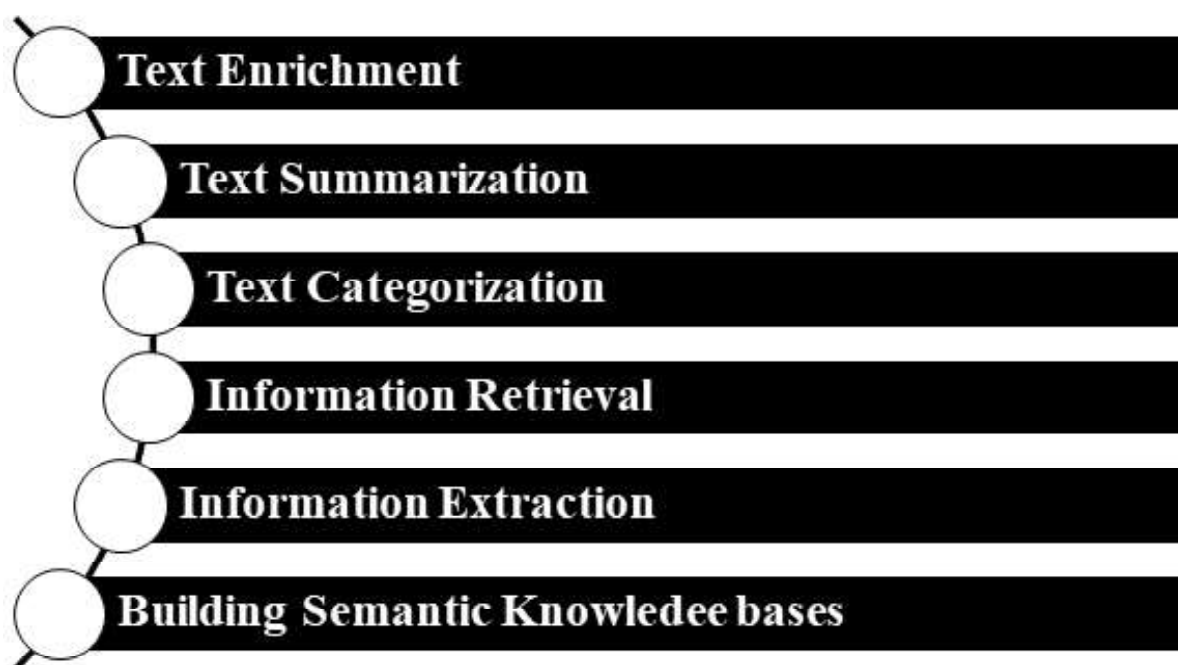


Figure 18 Different categories of text analytics tasks that have leveraged on Wikipedia's knowledge

6.1 Text Enrichment using Wikification

The natural language text available in the form of new items, journal article, course reading materials, blog entries or social media posts is associated with lot of ambiguity. While the degree of difficulty in reading and understanding such texts may vary depending on the reader, any form of background or secondary information about the text, made accessible contextually in a timely and non-intruding manner can make a big difference in helping the readers in comprehending the text.

When people have any difficulty in understanding a particular term/word in the text document, generally the user may search the meaning/definition of the difficult term in a search engine like Google. Further, it has been observed that in most cases the first page in the search engines result page is Wikipedia. Therefore, it would be a great help to the readers if the key terms in the text being read is automatically linked to the relevant pages on Wikipedia. This process is predominantly referred in the scholarly literature as Wikification.

Given an input document, the Wikification system identifies the important concepts in the text and automatically links these concepts to the corresponding Wikipedia pages (Mihalcea & Csomai, 2007).

The Wikification process involves multiple steps:

- ♦ Automatic keyword extraction.
- ♦ Word sense disambiguation.
- ♦ Linking the keywords to the relevant pages of wikipedia.

As the wikification or in other terms the Wikipedia style annotation of documents is very useful in a number of text processing tasks such as summarization, text categorization etc, the topic has been extensively researched. Different aspects have been studied by different researchers. While some works have focussed on the word sense disambiguation (Ratinov et al., 2011), which is a non-trivial task in wikification, other works

have focussed run-time to make the wikification process feasible for large data. Further, a lot of work is evident in the scholarly landscape about multilingual wikification.

There are a number of ways to wikify the text, which include online tools like <http://wikifier.org/> for quick wikification of small text, programming frameworks/APIs for integrating wikification in text applications (Getting Started with Python's Wikipedia API, n.d.) and browser extensions for enriching the web browsing experience. For example, the web page in figure 19 is the source for wikification and after the chrome plugin for wikification scanning is applied, all the keywords in the source are automatically linked to the relevant Wikipedia pages. Further, as depicted in figure 20, the snippet of the Wikipedia page is displayed in the form of popup when the mouse is over the keywords in the source page.

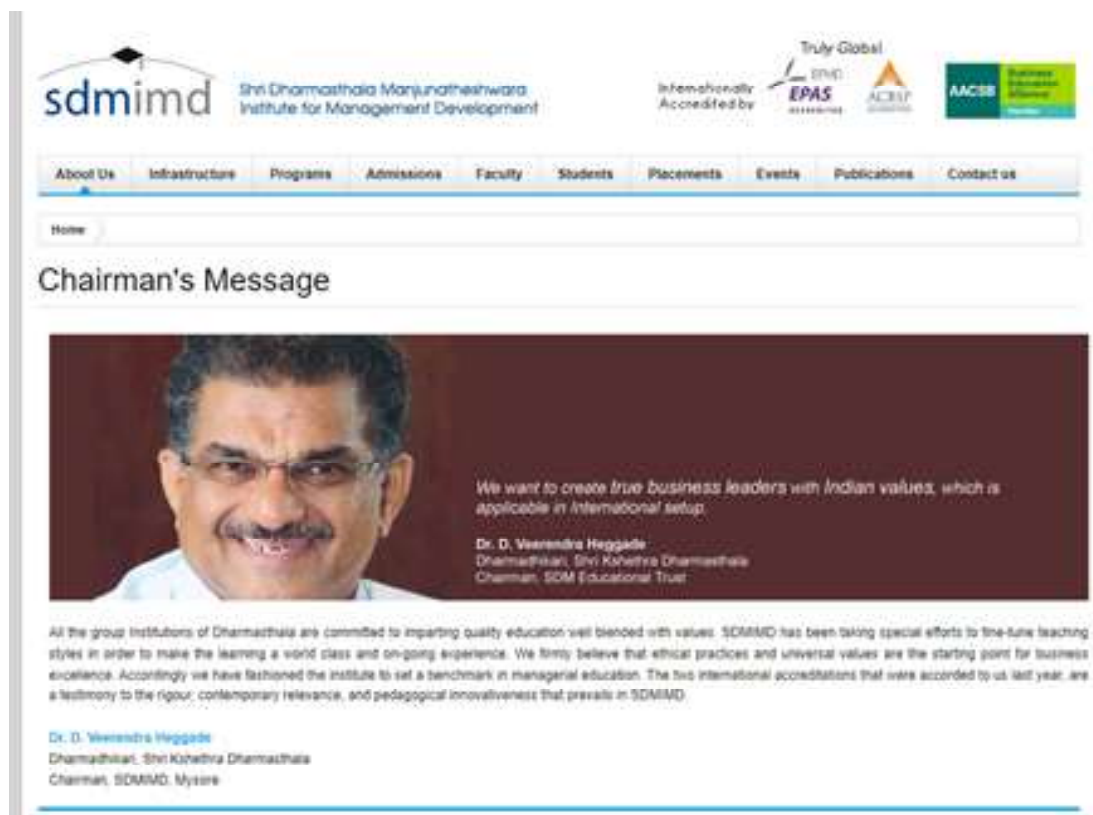


Figure 19 Sourc page for Wikification

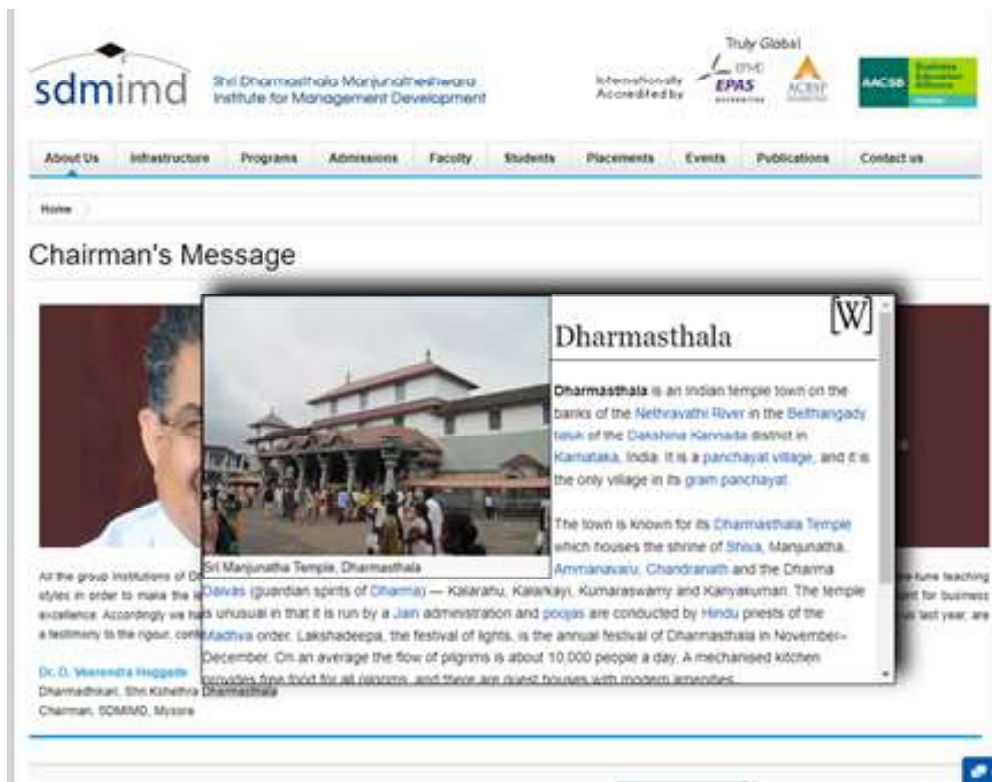


Figure 20 Page after Wikification, with wikilinks and auto- popup displaying the snippet of wiki page.

6.2 Text Summarization

Text summarization refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document. Automatic text summarization is an important task in machine learning and natural language processing (NLP).

There are broadly two approaches for Text Summarization:

- ♦ Extraction-based summarization
- ♦ Abstraction-based summarization

6.2.1 Extraction based summarization

The extractive text summarization technique involves pulling key phrases from the source document and combining them to make a summary. The extraction is made according to the defined metric without making any changes to the texts.

As a machine learning problem, extractive summarization entails weighing the essential sections of sentences and using the results to generate

summaries. Different types of algorithms and methods can be used to gauge the weights of the sentences and then rank them according to their relevance and similarity with one another for the purpose of joining them to generate a summary.

6.2.2 Abstraction based summarization

The abstraction technique entails paraphrasing and shortening parts of the source document.

Since abstractive machine learning algorithms can generate new phrases and sentences that represent the most important information from the source text, they can assist in overcoming the grammatical inaccuracies of the extraction techniques (A Gentle Introduction to Text Summarization in Machine Learning, 2019).

Several attempts have been made to leverage Wikipedia for text summarization. In the work by Ramanathan, Sankarasubramaniam, Mathur, and Gupta, the document sentences are mapped to

semantic concepts in Wikipedia and the sentences are selected for summary based on the frequency of the mapped concepts (Ramanathan et al., 2009). The work by Ye, Chua, and Lu is focussed on summarizing definitions using Wikipedia pages (*Summarizing definition from Wikipedia / Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, n.d.). One of the other prominent work is by Pourvali & Abadeh which leverage Wikipedia to form multiple independent graphs, and then use graph importance and lexical cohesion features for summarization (Pourvali & Abadeh, 2012).

6.3 Text Categorization

Text classification also known as text categorization or text tagging involves analysing a piece of text and assigning a set of predefined categories to it. Text classifiers can be used to organize, structure, and

categorize any type of textual data. For example, news articles can be organized by topics, support tickets can be organized by urgency, chat conversations can be organized by language, brand reviews can be organized by sentiment, and so on.

For example, if a customer review about a product is having the following text:

"The product has excellent features and is value for money"

A classifier can take this text as an input, analyse its content, and then automatically assign relevant tags, such as "Positive Review".

There are many approaches to automate the text classification, which are broadly grouped into the following types :

Ruled based	Rule-based approaches classify text into organized groups by using a set of handcrafted linguistic rules.
Machine Learning based	Instead of relying on manually crafted rules, text classification with machine learning learns to make classifications based on past observations. By using pre-labelled examples as training data, a machine learning algorithm can learn the different associations between pieces of text and that a particular output (i.e. tags) is expected for a particular input (i.e. text).
Hybrid Methods	Hybrids systems combine a base classifier trained with machine learning and a rule-based system, which is used to further improve the results. These hybrid systems can be easily fine-tuned by adding specific rules for those conflicting tags that haven't been correctly modeled by the base classifier.

Figure 21 Different approaches to Text Categorization

Many of the early works related to text classification were based on “Bag of Words” (BOW) representation, which only accounts for term frequency in the documents, and ignores important semantic relationships between key terms. To overcome this problem, some researchers have attempted to enrich text representation by means of manual intervention. However, considering the laborious and time-consuming process involved in manual enrichment of text, few researchers experimented the use of Wikipedia’s knowledge to automatically construct a thesaurus of concepts, which explicitly derives concept relationships based on the profuse structural knowledge of Wikipedia, including synonymy, polysemy, hyponymy, and associative relations (Wang et al., 2009). The generated thesaurus serves as a controlled vocabulary that bridges the variety of terminologies present in the corpus of documents. It facilitates the integration of the rich knowledge of Wikipedia into text documents, by resolving synonyms and introducing more general and associative concepts, which assist the identification of related topics among text documents. This in turn facilitates the classification of documents in a better way.

The other approach evident in the scholarly literature towards use of Wikipedia for text classification is by constructing Semantic Kernels (P. Wang & Domeniconi, 2008b). The key challenge involved in text classification is dealing with the sparsity and the high dimensionality of text data along with the complex semantics of the natural language. Although simple and commonly used, the BOW approach entails dealing with large dimensional data as each word or term in the textual corpus adds to the dimensionality of the table. Further, having only the words or terms of the corpus in the matrix, without the semantic relationship between the terms, the prediction capability of the classification algorithms reduces significantly. In their endeavour to improve the classification accuracy, Wang and Domeniconi, proposed embedding the background knowledge derived from Wikipedia into a semantic kernel which is then used to enrich the representation of documents. The empirical evaluation of their

approach with real data sets achieved improved classification accuracy compared to the conventional BoW approach.

The other prominent use case of Wikipedia has been classification of small text. With the widespread usage of internet and consequently many digital touch points, a lot of data about the internet users is generated in the form of web search snippets, forum, chat messages, customer review etc. When it comes to classifying such short texts, there is not enough word co-occurrence or shared context to achieve high accuracy. Although some pre-processing technologies such as removing stop words and stemming are proposed to improve the performance, normal machine learning methods usually fail to achieve desired accuracy due to the data sparseness and background knowledge. In their attempt to improve the accuracy of classification, Xiang Wang et al., mapped the short text to Wikipedia concepts and the concepts in turn were used to represent document for text categorization (X. Wang et al., 2013). After completing this process, use of traditional classification methods such as SVM, outperformed the traditional BoW approach.

6.4 Information Retrieval

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) (*Introduction to Information Retrieval*, n.d.). IR is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge discovery as a response to a user request (query). This process entails several steps starting with representing data and ending with returning relevant information to the user. The intermediate stage includes filtering, searching, matching and ranking operations. The primary goal of IR is to find the relevant information or a document that satisfies user’s information needs.

The Wikipedia’s knowledge base has been used in numerous ways to improve the IR process. The new IR

system developed by Liu which incorporated the word sense disambiguation algorithm and expanded queries using Wikipedia and WordNet dictionaries showed an increase in the performance of the retrieval in terms of recall, precision, mean and geometric mean average precisions(*Improve text retrieval effectiveness and robustness—ProQuest*, n.d.).

ESTER is another efficient search engine that works based on a combination of full text and ontology search(*ESTER / Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, n.d.). The search engine was built based on Wikipedia and YAGO ontology to process variety of complex queries in a fraction of second.

In a similar line of research, Vechtomova proposed new models for retrieving blog posts containing opinions about an entity expressed in the query by building a number of faceted queries (disjunctions of a list of short queries) using Wikipedia.

6.5 Information Extraction

Information Extraction is the task of automatically extracting information or facts from unstructured or semi-structured documents(Cowie & Lehnert, 1996). It usually serves as a starting point for other text mining algorithms(Allahyari et al., 2017). For example, extraction of entities and their relations from text can give us useful semantic information.

Information extraction is a technology based on analysing natural language in order to extract snippets of information. The process takes texts as input and produces fixed format, unambiguous data as output. This data may be used directly for display to users or may be stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in information retrieval applications such as internet search engine like Google(Davies et al., 2012).

The process of information extraction (IE) basically turns the unstructured information embedded in texts into structured data and typical sub-tasks of

information extraction include - named entity recognition, coreference resolution and relationship extraction (Liu, 2014).

While there are few studies about extracting information from Wikispedia itself (Ameta & Jat, 2018; Miháltz, 2010), many studies have focused on using Wikipedia's data for improving the IE tasks from other textual sources. It has been observed that several attempts have been made in the recent times to use Wikipedia for the purpose of extracting structured information from textual files like HTML, XML etc.

6.5.1 Named Entity Recognition

When it comes to Named Entity Recognition, which deals with identifying named entities such as personal names, names of organizations or genes from freeform text, generally Machine Learning and an annotated dictionary comes to play. Several researchers have used Wikipedia for NER. The team from the University of Economics, Prague, have developed NER systems based on Wikipedia's Search API as well as Apache Lucene search API(Sharma, 2018). Further, there are many popular frameworks for entity linking using Wikipedia, like Dexter, Babelfy etc. Dexter is an open source entity-linking framework developed by researchers at ISTI-CNR, Italy, Dexter identifies text fragments in a document referring to entities present in Wikipedia. Babelfy is a multilingual open source framework, which has a web interface and an API to perform entity-linking as well as word sense disambiguation.

6.5.2 Coreference Resolution

While extraction of entities from text is an important task for any type of analysis, another crucial aspect of text analytics is coreference resolution. Within a single document, coreference resolution finds the referents of expressions such as pronouns, demonstratives, or definite descriptions. On a collection of documents, cross-document coreference finds the sets of mentions for each distinct entity mentioned in the collection. Cross-document coreference is not only a useful output of information extraction in itself, but it also supports

Task	Description
Named Entity Recognition (NER)	Named entity recognition is the problem of finding references to entities (mentions) in natural language text, and labelling them with their location and type (Leaman & Gonzalez, 2008). What constitutes a named-entity type is application specific and commonly consist of people, places or organizations but also include specific entities such as names of genes and proteins(You et al., 2014).
Coreference Resolution	It involves detection of coreference and anaphoric links between text entities. Coreference resolution is the task of finding all expressions that refer to the same entity in a text. It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction(<i>The Stanford Natural Language Processing Group</i> , n.d.).
Relationship Extraction	Relation Extraction examines pairs of entities in a document to determine whether a relation exists between them (Culotta et al., 2006; Konstantinova, 2014). It involves identification of relations between entities, such as PERSON works for ORGANIZATION (extracted from the sentence "X works for Microsoft."). A relationship extraction task requires the detection and classification of semantic relationship that exist within a set of entities

Table 3 Typical sub-tasks of Information Extraction

other information extraction tasks. The coreference resolution also plays a pivotal role in knowledge base construction and is very useful for joint inference with other NLP components.

The training and testing of cross document coreference requires a large labeled dataset and obtaining such large scale organic labeled dataset is very difficult. One prominent solution to this problem is Wikilinks.

6.5.3 Relation Extraction:

While the extraction of entities like persons and organizations is important because they form the most basic unit of the information, it is the extraction of relations between those entities that play a key role in natural language processing and better understanding of text. Relation extraction involves identifying the links between named entities and deciding which ones are meaningful for the concrete application or problem. Given two entities, the aim is at locating the occurrence of a specific relationship type between them.

A relation usually denotes a well - defined relationship between two or more Named- Entities. Most of the traditional methods are feature based and treat this task as a pipeline of two separate tasks, i.e., Named Entity Recognition and Relation Classification.

Name entity recognition is a challenging task which needs massive prior knowledge sources for better performance. The research community has suggested several approaches and have employed them on different domains. Early works have focused on heuristic and handcrafted rules. By defining the formation patterns and context over lexical-syntactic features and term constituents, entities are recognized by matching the patterns against the input documents. The rule-based system may achieve high degree of precision. However, the development process is time-consuming and porting these developed rules from one domain to another is a major challenge. Realizing the constraints of rule-based approach, most recent research in Name Entity Recognition tends to use machine learning approach. The learning methods include various supervised, semi-supervised and unsupervised processes. The supervised learning tends to be the dominant technique for named entity recognition and classification. However, supervised machine learning methods require large number of annotated documents for model training and its performance typically depends on the availability of sufficient high-quality training data in the domain of interest. Further, there are some systems which have employed hybrid methods to combine different rule-based and/or machine learning systems for improved performance over individual approaches.

In the pipeline approach, the different techniques mentioned above are used to first extract the entities before embarking on the task of identifying the relations between entities. Such an approach assumes that the knowledge about boundaries and types of entity mentions are known beforehand. If such knowledge is not available, one needs to apply some technique to identify the entities and then apply appropriate relation extraction technique. However, such a pipeline method is prone to propagation of

errors from one phase to another. To avoid this propagation of errors, there is a line of research which extracts entities and relations jointly. Considering that both entity and relation extraction can benefit from being performed jointly, allowing each task to correct the errors of the other.

Wikipedia as a knowledge base has been used by many researchers in relation extractions tasks. One of the prominent works in this direction is minimally-supervised extraction of domain-specific part-whole relations using Wikipedia as knowledge-base by Ashwin Ittoo and Gosse Bouma. The crux of their approach lies in applying a minimally-supervised algorithm to a large, broad-coverage corpus, which they used as knowledge-base. From this knowledge-base, a set of patterns were acquired that reliably express part-whole relations. Further, all triples consisting of the acquired patterns and the instance pairs they connect, were extracted from a domain-specific text collection. These triples established the domain-specific part-whole relations.

The other line of research evident in the recent scholarly literature pertaining to Relation Extraction using Wikipedia is the use of Distant Supervision. The focus of Nguyen et.al work is end-to-end relation extraction using distant supervision from external semantic repositories like Wikipedia and YAGO.

The other recent use-case of Wikipedia is Open Information Extraction. While Machine Learning has become the most prominent method for any IE tasks like Relation Extraction, it requires identifying target relations ahead of time and involves the laborious construction of a labelled training set. As a result, supervised learning techniques cannot scale to the large data sets. This lacuna has triggered a different approach, referred as Open Information Extraction and involves a task of extracting information to fill an unbounded number of relational schemata, whose structure is unknown in advance. To bootstrap such a process, Wikipedia is used(Weld et al., 2009).

6.6 Building Semantic Knowledge Bases

Wikipedia has been the most successful collaborative encyclopedia and is among the widely used websites around the globe. Wikipedia's English edition alone has around 6 million articles. However, the limited search mechanism provided on Wikipedia and the form in which its data is stored, brings in many limitations for the use of Wikipedia by users and direct interpretation by machines. In spite of these lacunae, Wikipedia continues to entice the research community because its data is not only comprehensive but has well-formed structure and hierarchical categorization as well. To effectively use the knowledge that is concealed in Wikipedia, several attempts have been made to extract and transform the unstructured and semi-structured data of Wikipedia, into structured and semantically enriched knowledge bases.

A knowledge base (KB) is a technology used to store complex structured and unstructured information used by a computer system. The initial use of the term was in connection with expert systems which were the first knowledge-based systems.

In recent years, several noteworthy large, cross-domain, and openly available knowledge bases have been created. These include DBpedia, Wikidata, YAGO etc. (Färber et al., 2017). These knowledge bases are not only enabling effective use of the Wikipedia's concealed knowledge by humans, they are facilitating better and faster interpretation of knowledge by machines as well. This has resulted in development of many smart applications including Question-Answering Systems.

The Knowledge Bases built using Wikipedia, have employed different approaches for data curation and storage. The retrieval mechanisms facilitated by these Knowledge Bases are also different. Further, they also differ in their dept and breadth of knowledge.

6.6.1 DBpedia

The primary objective of DBpedia project was to convert Wikipedia content into structured knowledge,

which could facilitate the use of Semantic Web techniques, such as asking sophisticated queries against Wikipedia, linking it to other datasets on the Web, or creating new applications or mashups. The project in its original form developed an information extraction framework to convert Wikipedia content to RDF and consisted of 103 million RDF triples. Besides, developing a web interface to access the knowledge base, the open and linked nature of the DBpedia facilitated in linking its content to other open knowledge bases and integration with other semantic technologies. After its initial version developed in 2007, the DBpedia project has constantly improved and extended by a large global community. DBpedia has been the precursor of today's knowledge graphs, driving prototyping, proof-of-concepts and innovation. Many enterprises, such as Apple (via Siri), Google (via Freebase and Google Knowledge Graph), and IBM (via Watson), and particularly their respective high-visibility projects associated with artificial intelligence, have adopted the idea of data extraction from Wikipedia.

6.6.2 YAGO

The motivation behind the development of YAGO was development of a huge ontology with knowledge from several sources, instead of relying on single source of background knowledge. While the core of YAGO was assembled from Wikipedia, but rather than using information extraction methods to leverage the knowledge of Wikipedia, YAGO utilizes the category pages of Wikipedia. Category pages which have lists of articles that belong to a specific category were used to get the candidates for entities, concepts and relations. One of the key requirements of any ontology is arrangement of concepts in a taxonomy. Though Wikipedia categories are arranged in a hierarchy but are not very useful for ontological purposes. For example, in Wikipedia, the information about a football player X, who is a citizen of country C, may have super category named "Football in C", wherein X may be interpreted as a Football and not as a player. WordNet, in contrast, provides a clean and carefully assembled hierarchy of thousands of concepts. But

the Wikipedia concepts have no obvious counterparts in WordNet. YAGO project developed a technique that link two sources with near-perfect accuracy. This approach enabled YAGO to use the vast number of individuals known in Wikipedia, coupled with exploitation of clean taxonomy of concepts from WordNet. Since its initial release in 2008, YAGO has undergone several improvements and because of the ontology's high coverage and high quality, it has been employed for several AI systems. A prominent application of YAGO, is its use in the IBM Watson.

6.6.3 Wikidata

Wikimedia Foundation started Wikidata project with an objective of creating a central storage for the

structured data of its sister projects including Wikipedia. With the collaborative and multilingual nature of Wikipedia, the same information often appears in articles in many languages and on many articles within a single language. Population numbers for Rome, for example, can be found in the English and Italian article about Rome, but also in the English article Cities in Italy. All these numbers are different. The goal of Wikidata is to overcome these problems by creating new ways for Wikipedia to manage its data on a global scale. Wikidata was launched in 2012 and today it is one of the widely used, free and open knowledge bases that can be read and edited by both humans and machines.

Table 4 Key information about KBs built using Wikipedia

	DBpedia	YAGO	Wikidata
Developed by	Leipzig University University of Mannheim OpenLink	Max-Planck-Institute Saarbrücken	Wikimedia Foundation
Website	dbpedia.org	https://www.mpi-inf.mpg.de/yago-naga/yago/	www.wikidata.org
Launched	2007	2008	2012
License	GNU General Public License	Creative Commons	Public Domain License
Sources of data	Wikipedia	Wikipedia and WordNet	Wikipedia

7 Conclusions and Future Work

While the use of Wikipedia for scholarly activities is a subject of debate, Wikipedia is among the top 10 websites in terms of worldwide popularity / usage (Alexa). it is being used extensively by all types of web users, as an easily accessible tertiary source of information about anything and everything, as a quick "ready reference", to get a sense of a concept or idea ("Wikipedia," 2019). However, the limited search mechanism provided on Wikipedia and the form in which its data is stored, brings in many limitations for the use of Wikipedia by users and direct interpretation by machines. Despite these lacunae, Wikipedia

continues to entice the research community because its data is not only comprehensive but has well-formed structure and hierarchical categorization as well.

The current exploratory study found that the collaborative knowledge base of Wikipedia has been used in the recent years extensively for different processes and applications related to Text Analytics. The wiki annotations or wikification has been used to aid the text enrichment and facilitate better understanding of text. When used as the background information, it was found that Wikipedia can play a pivotal role in summarization of text documents. The content organisation and rich semantic elements of

Wikipedia have helped in several studies to improve significantly the Information Retrieval and Information Extraction tasks. The open nature of Wikipedia and its constantly updated well-formed information has been used in the creation of ontologies and consequently has backed many contemporary knowledge systems such as automated question - answering, speech recognition etc.

With the increase in the volume and velocity of text data in organisations along with better computing facilities, there has been a paradigm shift in the text analytics approach. Faster results with least manual intervention have become a norm and consequently Machine Learning (ML) has taken a centre stage. The biggest requirement for any ML activity is access to large background data for the purpose of training, testing and validating the ML models. Wikipedia can be fitting solution for this requirement of ML as well. Considering the fitness of Wikipedia for ML tasks, most analytical and ML software tools, programming frameworks are provided direct support for leveraging the Wikipedia content in the form of plugins/widgets, APIs and programming libraries. Further, it has been observed that Wikipedia Developers and the user/developer community have also developed many tools and technologies which can be employed for different text analytics applications.

As far as the future scope of this study is concerned, besides further developments related to semantic technologies for using Wikipedia in smarter knowledge applications, the other key theme found to be promising for further research is the use of Wikipedia for ML, particularly for distant machine learning.

8 References

- A Gentle Introduction to Text Summarization in Machine Learning. (2019, April 16). FloydHub Blog. <https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/>
- A Survey of Information Retrieval and Filtering Methods. (n.d.). Retrieved September 26, 2018, from <https://drum.lib.umd.edu/handle/1903/436>
- About Cochrane Reviews | Cochrane Library. (n.d.). Retrieved November 3, 2018, from <https://www.cochranelibrary.com/about/about-cochrane-reviews>
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. ArXiv:1707.02919 [Cs]. <http://arxiv.org/abs/1707.02919>
- Ameta, D., & Jat, P. M. (2018). Information extraction from wikipedia articles using DeepDive. 2018 International Conference on Communication Information and Computing Technology (ICCICT), 1-6. <https://doi.org/10.1109/ICCICT.2018.8325869>
- An introduction to Machine Learning. (2017, August 24). GeeksforGeeks. <https://www.geeksforgeeks.org/introduction-machine-learning/>
- Banerjee, S., Ramanathan, K., & Gupta, A. (2007). Clustering short texts using wikipedia. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 787-788.
- Buscaldi, D., & Rosso, P. (2006). Mining knowledge from wikipedia for the question answering task. Proceedings of the International Conference on Language Resources and Evaluation, 727-730.
- Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), 1165-1188. JSTOR. <https://doi.org/10.2307/41703503>
- Cowie, J., & Lehnert, W. (1996). Information Extraction. Commun. ACM, 39(1), 80-91. <https://doi.org/10.1145/234173.234209>
- Culotta, A., McCallum, A., & Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 296-303.
- Davies, J., Studer, R., & Warren, P. (2012). Semantic Web Technologies. Wiley.
- ESTER | Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. (n.d.). Retrieved March 31, 2020, from <https://dl.acm.org/doi/abs/10.1145/1277741.1277856>

- Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2017). Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1), 77-129. <https://doi.org/10.3233/SW-170275>
- Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 112-117. <http://dl.acm.org/citation.cfm?id=3001335.3001354>
- Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). Discovering context: Classifying tweets through a semantic transform based on wikipedia. *International Conference on Foundations of Augmented Cognition*, 484-492.
- Getting Started with Python's Wikipedia API. (n.d.). Stack Abuse. Retrieved April 2, 2020, from <https://stackabuse.com/getting-started-with-pythons-wikipedia-api/>
- Haneem, F., Ali, R., Kama, N., & Basri, S. (2017). Descriptive analysis and text analysis in Systematic Literature Review: A review of Master Data Management. 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), 1-6. <https://doi.org/10.1109/ICRIIS.2017.8002473>
- Hotho, A., Nurnberger, A., Paaß, G., & Augustin, S. (n.d.). A Brief Survey of Text Mining. 37.
- Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (2009). Exploiting Wikipedia as external knowledge for document clustering. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 389-396.
- Improve text retrieval effectiveness and robustness-ProQuest. (n.d.). Retrieved March 31, 2020, from <https://search.proquest.com/docview/304950228>
- Introduction to Information Retrieval. (n.d.). Retrieved March 31, 2020, from <https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html>
- Konstantinova, N. (2014). Review of relation extraction methods: What is new out there? *International Conference on Analysis of Images, Social Networks and Texts_x000D_*, 15-28.
- Leaman, R., & Gonzalez, G. (2008). BANNER: An executable survey of advances in biomedical named entity recognition. In *Biocomputing 2008* (pp. 652-663). World Scientific.
- Liddy, E. D. (n.d.). *Natural Language Processing*. 15.
- Liu, X. (2014). Fast recursive biomedical event extraction [PhD Thesis]. Université de Technologie de Compiègne.
- Mihalcea, R., & Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 233-242.
- Miháltz, M. (2010). Information Extraction from Wikipedia Using Pattern Learning. *Acta Cybern.*, 19, 677-694.
- Mohler, T. (2019, September 9). The 7 Basic Functions of Text Analytics. Lexalytics. <https://www.lexalytics.com/lexablog/text-analytics-functions-explained>
- Pourvali, M., & Abadeh, P. D. M. S. (2012). A new graph based text segmentation using Wikipedia for automatic text summarization. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 3(1).
- Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the Special Issue on Summarization. *Computational Linguistics*, 28(4), 399-408. <https://doi.org/10.1162/089120102762671927>
- Ramanathan, K., Sankarasubramaniam, Y., Mathur, N., & Gupta, A. (2009). Document summarization using Wikipedia. *Proceedings of the First International Conference on Intelligent Human Computer Interaction*, 254-260.
- Ratinov, L., Roth, D., Downey, D., & Anderson, M. (2011). Local and Global Algorithms for Disambiguation to Wikipedia. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1375-1384. <https://www.aclweb.org/anthology/P11-1138>
- Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization of the World from Edge to Core (IDC White Paper - #US44413318; p. 28). <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-data-age-whitepaper.pdf>
- Ryu, P.-M., Jang, M.-G., & Kim, H.-K. (2014). Open domain question answering using Wikipedia-based knowledge model. *Information Processing & Management*, 50(5), 683-692.

- Sarkar, D. (2016). Text Analytics with Python. In Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data. Apress.
- Sharma, A. (2018, August 24). Exploring Named-Entity Recognition With Wikipedia. Analytics India Magazine. <https://analyticsindiamag.com/exploring-named-entity-recognition-with-wikipedia/>
- Štrukelj, E. (2018, January 3). Writing a Systematic Literature Review. JEPS Bulletin. <https://blog.efpsa.org/2018/01/03/writing-a-systematic-literature-review/>
- Summarizing definition from Wikipedia | Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1 (world). (n.d.). Retrieved April 2, 2020, from <https://dl.acm.org/doi/abs/10.5555/1687878.1687908>
- The biggest data challenges that you might not even know you have-Watson. (n.d.). Retrieved March 19, 2020, from <https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/>
- The Stanford Natural Language Processing Group. (n.d.). Retrieved November 12, 2018, from <https://nlp.stanford.edu/projects/coref.shtml>
- The teacher's guide to Creative Commons licenses | Open Education Europa. (n.d.). Retrieved March 13, 2020, from <https://web.archive.org/web/20180626111219/http://www.openeducationeuropa.eu/en/blogs/teachers-guide-creative-commons-licenses>
- The technology behind free knowledge. (2019, May 3). Wikimedia Foundation. <https://wikimediafoundation.org/about/2018-annual-report/technology/>
- The Use of Google Scholar for Research and Research Dissemination. (n.d.). Retrieved November 4, 2018, from <https://onlinelibrary.wiley.com/doi/full/10.1002/nha3.20209#nha320209-bib-0008>
- Underhill, D. G., McDowell, L. K., Marchette, D. J., & Solka, J. L. (2007). Enhancing Text Analysis via Dimensionality Reduction. 2007 IEEE International Conference on Information Reuse and Integration, 348-353. <https://doi.org/10.1109/IRI.2007.4296645>
- Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50(1), 104-112. <https://doi.org/10.1016/j.ipm.2013.08.006>
- Wang, P., & Domeniconi, C. (2008a). Building semantic kernels for text classification using wikipedia. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 713-721.
- Wang, P., & Domeniconi, C. (2008b). Building semantic kernels for text classification using wikipedia. Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08, 713. <https://doi.org/10.1145/1401890.1401976>
- Wang, P., Hu, J., Zeng, H.-J., & Chen, Z. (2009a). Using Wikipedia knowledge to improve text classification. Knowledge and Information Systems, 19(3), 265-281.
- Wang, P., Hu, J., Zeng, H.-J., & Chen, Z. (2009b). Using Wikipedia knowledge to improve text classification. Knowledge and Information Systems, 19(3), 265-281. <https://doi.org/10.1007/s10115-008-0152-4>
- Wang, X., Chen, R., Jia, Y., & Zhou, B. (2013). Short Text Classification Using Wikipedia Concept Based Document Representation. 2013 International Conference on Information Technology and Applications, 471-474. <https://doi.org/10.1109/ITA.2013.114>
- Weld, D. S., Hoffmann, R., & Wu, F. (2008). Using wikipedia to bootstrap open information extraction. SIGMOD Record, 37(4), 62-68.
- Weld, D. S., Hoffmann, R., & Wu, F. (2009). Using wikipedia to bootstrap open information extraction. Acm Sigmod Record, 37(4), 62-68.
- What is Text Classification? (n.d.). MonkeyLearn. Retrieved March 24, 2020, from <https://monkeylearn.com/what-is-text-classification>
- Wickelmaier, F. (2003). An introduction to MDS. Sound Quality Research Unit, Aalborg University, Denmark, 46(5), 1-26.
- Wiki. (2020). In Wikipedia. <https://en.wikipedia.org/w/index.php?title=Wiki&oldid=942581647>
- Wikipedia:About. (2019a). In Wikipedia. <https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=926297166>
- Wikipedia:Academic use. (2019b). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Academic_use&oldid=921100309

- Wikipedia:Contents/Overviews. (2019c). In Wikipedia. <https://en.wikipedia.org/w/index.php?title=Wikipedia:Contents/Overviews&oldid=926544356>
- Wikipedia:FAQ/Technical. (2020). In Wikipedia. <https://en.wikipedia.org/w/index.php?title=Wikipedia:FAQ/Technical&oldid=941911136>
- Wikipedia's coverage of essential vaccines is expanding. (2016, March 9). Wikimedia Foundation. <https://wikimediafoundation.org/news/2016/03/09/wikipedias-essential-vaccines/>
- Wu, F., & Weld, D. S. (2010). Open information extraction using Wikipedia. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 118-127.
- You, D., Antani, S., Demner-Fushman, D., & Thoma, G. R. (2014). Biomedical image segmentation for semantic visual feature extraction. Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference On, 289-292.