REVA UNIVERSITY
Bengaluru, India

A Project Report on

Machine Learning approach to Text Summarization for Meta Descriptions

Submitted in partial fulfilment for the award of the degree of

Master in Business Administration

In **Business Analytics**

Submitted by

Akanksha Prasad
R18DM002

Under the Guidance of

Dr. JB Simha
Chief Technology Officer, ABIBA Systems

REVA University
Rukmini Knowledge Park, Kattigenahalli,
Yelahanka, Bangalore – 560064
September, 2020

REVA
UNIVERSITY
Bengaluru, India

Candidate’s Declaration

I, Akanksha Prasad, hereby declare that I have finished the first project work towards Master in Business Administration course at, REVA University on the topic titled ‘**An approach to building machine-generated search-friendly meta descriptions for websites’** under the supervision of Dr. JB Simha, Chief Technology Officer, ABIBA Systems. This report embodies the creative work done by me in partial fulfilment of the requirements for the award of degree for the academic year of 2020.

Place: Bengaluru Name of the Student: Akanksha
Prasad
Date: October , 2020 Signature of Student

akanksha

REVA
UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the PROJECT work titled '**Machine Learning approach to Text Summarization for Meta Descriptions'** is carried out by Akanksha Prasad with SRN R18DM002 is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of Master's in Business Administration in Business Analytics during the academic year 2020. The Project report has been tested for plagiarism and has passed the plagiarism test with the similarity score of less than 15%. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.

<Signature of the Guide>
Dr. JB Simha, CTO, ABIBA Systems
Guide

<Signature of the Director>
<Name of the Director>
Director

External Viva
Names of the Examiners

1. <Name> <Designation> <Signature>
2. <Name> <Designation> <Signature>

Place: Bengaluru
Date: October, 2020

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	NLTK	Natural Language Toolkit
2	NLP	Natural language processing
3	SERP	Search Engine Rank Page
4	SEO	Search Engine Optimization
5	TF-IDF	term frequency–inverse document frequency

List of Figures

Abstract

Meta descriptions are the short snippets about the website present in the search result. These descriptions are trivial in bringing prospects to the website and ensuring conversion. However trivial in nature, meta descriptions are often ignored by businesses and the result is most of the websites either do not have any description or the descriptions are inappropriate.

Another challenge is generating the descriptions and turnaround time associated with it, therefore, the businesses end up with inappropriate descriptions and they witness lower conversions or click through from search engines.

An approach to building machine-generation description addresses the challenge of needing extra resource and the longer turnaround time. In this paper, we try a unique approach of rule-based supervised learning by leveraging machine learning to build meta descriptions for the website and we pull out the meta description from leading and popular websites in similar space to build a corpus of the relevant, search-friendly description of the desired quality.

Keywords: Text Summarization, Text Analytics, Natural Language Processing, Meta Description, SEO, Google search, SERP, Rankbrain, Semantics,

Contents

Chapter 1: Introduction

Early 2000 saw the advent of dot.com era. It encouraged business towards setting up their online presence and building websites. Since then every new business started with a website and being visible on search engines like Google for growing visibility in the market. Some reports indicate that today there are more than 17 billion websites from India and this number will continue to grow year-after-year.

Over the last 20 years, it has been established that online presence helps in boosting brand awareness and even brining more qualified leads for sales. A study conducted using 156 customers with online presence revealed that a relative correlation between having an online presence with a quality website and higher business performance.(Lee & Kozar, 2006)

Being present online was not just about being visible, but the concept of building online presence was also related to discovery. Discovery is when a prospect customer, looks up on search engine for a solution and comes across the website of respective business. The founders of Google in their paper (Page & Brin, 1998) elaborated this concept and methodology of search and discovery with the help of hypertext and keywords. In the paper, Google was introduced as a prototype of the search engine for large-scale use, which used the hypertext for identification. The algorithm in Google was designed to inspect and index the internet and produce satisfying results of the search.

In the study, they used a database of close to 20 mn pages and engineered the engine. Terms and keywords played a role in helping index and discover millions of websites. Moving aside from the traditional search and discovery techniques used at such large scale, the paper attempted to leverage hypertext to produce search results. It also built a new way for a large-scale system that exploited the information in hypertext and helped in discovery of particular websites.

In the last 20 years, Google introduced many improvements and newer versions of its search algorithm. All of these major algorithms aimed at making search and discovery better and closer to the expected results.

To bring out most useful data, these search algorithms looked at many factors, like query words, relevance of pages, importance of sources, the relevance with location and more. Digital marketers have been working on Search engine rank page (SERP) optimization. While all the Search Engine Optimization (SEO) tactics, help in discovery, they do not promise conversion from discovery.

The definition of SEO often includes specifications for increased traffic to a given website, improved quality of traffic, increased profits, or brand awareness. Any search query displays results on the page. Most search queries have 10 results per page, and up to 10 search pages. Google sorts through billions of websites in the Search index to find and give out the most relevant and useful results in search.

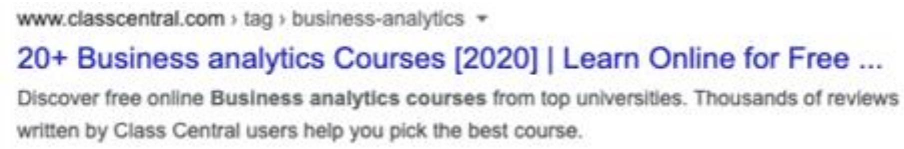
But the audiences looking are interested in the top few results only. So majority of the digital marketeers are focused at being at the top to be discoverable in the top few results. (Swati et al., 2013) The results appearing in the first page receive 95% of the clicks, and the numbers drop up to 3 % up to the 5th page. Search results from 6-10 pages are rarely to almost never look at with less than 1% of the visitors searching beyond 5th page or 50th search result.

The SEO tactics, help in improving the Search Engine Result Page (SERP), making a website appear within top few search results, but they do not guarantee the conversion and whether the audience would click the respective result and enter the website to get more details.

Search Engine optimization is a constant pursuit and takes consistent effort to improve page ranking in the search results. After so much effort and years of build-up, few out of thousands of websites make it to the top 50 search results. However, a wrong meta description can damage the entire effort.

Meta description play an important in bringing conversion. A meta description is used in the search result as a part of snippet shown to the visitors in search query results. A good meta description should generally inform the users about what a page is about and what they should be expecting and looking for, with a relevant and short summary (Raj Krishnan, 2007). Absence of the meta descriptions of the pages in search results, fails to provide website information and ends up without. While, good-quality descriptions give users very clear idea of website content. There have been some study around understanding how Google's search engine works, find the variations in the meta description for various sites in different languages.

It has been noticed that a lot of business, despite putting tremendous efforts in SEO are unable to gain a lot of website presence because of inappropriate meta descriptions. Wrong descriptions mean website visitors do not click the website as the description is misleading or not attractive enough, and for the limited website traffic that comes, the users exit because the information is not relevant.



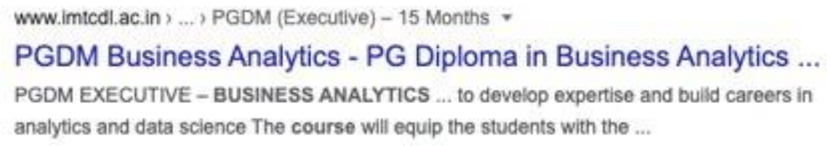
www.classcentral.com › tag › business-analytics ▾

20+ Business analytics Courses [2020] | Learn Online for Free ...

Discover free online **Business analytics courses** from top universities. Thousands of reviews written by Class Central users help you pick the best course.

Figure No. 1(a)

The figure above, is an example of a typical search result with headline, link and the meta description of the respective webpage. It helps highlight what a reader would expect in the webpage and the benefits of visiting the website.



www.imtcdl.ac.in › ... › PGDM (Executive) – 15 Months ▾

PGDM Business Analytics - PG Diploma in Business Analytics ...

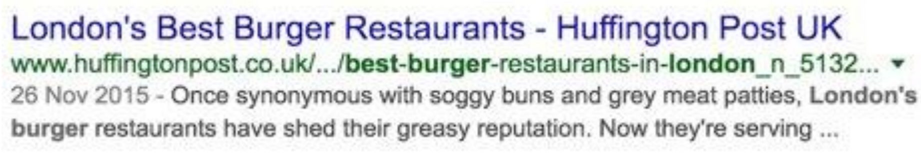
PGDM EXECUTIVE – **BUSINESS ANALYTICS** ... to develop expertise and build careers in analytics and data science The **course** will equip the students with the ...

Figure No. 1(b)

The figure above is an example of inappropriate meta description, which has been randomly picked by Google from the web content, and therefore it reflects sentences ending abruptly.

The reason why many businesses get meta description wrong is because of the lack of resource, lack of involvement of business SME and also because the web development and we management is largely managed by technical resources, who may/may not have understanding of the business.

At such scenario, the web developers end up selecting the first sentence as the summary of the page, which may be not correct.



London's Best Burger Restaurants - Huffington Post UK

www.huffingtonpost.co.uk/.../best-burger-restaurants-in-london_n_5132...

26 Nov 2015 - Once synonymous with soggy buns and grey meat patties, London's burger restaurants have shed their greasy reputation. Now they're serving ...

Figure No. 1(c)

In the figure captured above, the link is from Huffington post and it talks about an article about best burger restaurants in London. But the description instead of giving a view of the article stars with a story…Once upon a time.

Some of the common mistakes with meta descriptions are

- Misleading description
- Unimpressive content of description
- Irrelevant information
- Website content and description mismatch
- And longer than required description, which is truncated to half meaningless sentences

Wrong or inappropriate meta descriptions leads to loss of customers interest in the website. Either the website conversations are low or the bounce rate and exit rates are high. People leave within seconds of entering the webpage.

The expected business outcome of every marketing and digital marketing effort is increasing leads and hence more business. And higher bounce rates and reduced conversion rates mean no leads and no business. Furthermore, a typical website has over 20 links, each link is a webpage in itself, needing a unique meta description. Larger the organization, higher the number of web pages and required descriptions. And as the organization grows, its websites changes and so should its meta description.

While the businesses that get right at meta descriptions, they spend a lot of time in working internally and coordinating between various departments to get the right description. While being a time-consuming process, this also becomes a costly effort, needing a content and domain specialist.

This paper is an attempt at building an approach to solving this business challenge and finding a way to build machine-generated search-friendly meta descriptions for websites with the help of text summarization approach using machine learning techniques. There are a set of research in the area to see the current progress.

Chapter 2: Literature Review

A majority of the study in the field of search has been around addressing the challenge of discoverability. For many businesses, search engine optimization is purely around improving the ranking in the search results and appearing in the first few pages, because as described above, visitors' visits start reducing after 3-5 pages.

Search Engine Optimization is working on various on-page and off-page tactics to improve page rank. It includes working on the keywords in the page, tagging the page assets for searchability, adopting other best practices.(Schröer & Schröer, 2018) This has been an industry question and an active field of study.

Some reports explore various ways to achieve better ranking on search results. They have proposed a set of guidelines by leveraging backlinks, or specific keywords through keyword research and by building the referrals. (Khan & Mahmood, 2018)

While some reports propose an optimization mechanism that could be used by exploring the internet marketing or social media marketing approach to enhance the visibility and exposure of the website, thus improving the ranking in search results..(Shih et al., 2013)

A wider scope of research is trying to explore the ways to bring in the original intent of search into the results and make the results more accurate. But the quality of searches rarely matches up. In this blog, (Max Irwin, 2018) the author elucidates that the relevance of search is every so often equated with considered that prospects have found what they were looking for. There is the question around the relevance of search and the experience overall. If visitors do not move beyond the first page, they would have gotten what they needed. When considering enhancements to search, it is essential to consider the overall quality.

Google reached out to what it defines as “Search Quality Raters” or “Raters” to help it evaluate the quality of search results. These Raters evaluate the quality of pages based on more than 20 criteria including content accuracy and content relevance. (“Google: Search Quality Raters,” 2020)

The issue of search quality also comes into questions when you look for the descriptions and you do not find what you get, because either the search is not relevant and appropriate or the search qualifier, the meta description is inappropriate.

Creating a good meta description needs good content and domain knowledge. It is the first impression of the website that any visitor gets. And creating a best info of the website in such smaller space needs a good idea of the messaging that the business wants to put across.

A good meta description should not be longer than 150 characters or 20-30 words. There has been some study around understanding how Google’s search engine works, find the variations in the meta description for various sites in different languages. The study captured good relations and patterns in meta descriptions of certain websites. Like pages in Western European languages had longer descriptions than the Chinese pages. (Craven, 2004)

In another paper, in understanding and working on improvising meta descriptions, the author has looked into regular phrases, the syntactic structure of sentences and the content. The report deep dives into the density of words in description and density of bi-gram description phrases. And finds the results with metatag descriptions showed better results in densities.

When it comes to building machines –generated description, the first step is connecting the world of literature with the world of binary. This working on the semantics of words.

A look into the academic work around semantics suggest, researchers have been working on methodologies on making sentences more comprehensible to computers. There has been a study on Vector space models (VSMs) and semantic processing of text. In this, computers are made to understand the meaning of our language. There are limits to our ability to give instructions to computers, and for the computers to explain their actions back to us.(Salton et al., 1996) This also includes making computers to analyse and process words. Vector space models (VSMs) for semantics is a beginning to address the existing limits. This paper explores the use of VSMs for semantic processing by organizing the VSMs according to a structure of a matrix in a model.

Another area of research is in searching and locating content in multimedia sites. In a paper, proposed a platform for the development of multimedia web information systems. With an approach based on the combination of semantic web technologies and collaborative tagging. Producers can add meta-data to multimedia content associating it with different domain-specific ontologies. At the same time, users can tag the content in a collaborative way. (Labra Gayo et al., 2010)

In 2015, Google used the underlying technique of making works understandable to the machines. First introduced in 2015, today Rankbrain is an important part of the search algorithm and this gives us a good insight into understanding how to leverage machines for working around words. RankBrain takes advantage of artificial intelligence to install large amounts of written words into mathematical items as vectors. These are understandable by the computer. When RankBrain notices unfamiliar words or phrases, the machine makes a theory to what it could mean or be connected to using certain filters. This makes it very effective at managing unfamiliar search queries. (Danny Sullivan, 2015)

Earlier, Google might search for web pages with exact search keyword. But lately, and especially with the introduction of the Hummingbird algorithm in the searches, the results on Google have improved at recognizing connections between words.

But aside the breakthrough from Google there is limited research done in particular in the area of using and optimizing meta descriptions. The priority of improving search result and search engine optimization has often taken precedence over tips on inproving meta description. This can be explained because while meta description ensures conversion, SEO ensures discoverability, which is the first and the most important step in the sales funnel

Chapter 3: Problem Statement

Businesses invest time and effort is setting up the best website that could bring customers to the webpages and convince them to buy product or service. All of the digital marketing efforts go in getting visible on search engines for the target audience. The web pages and website, management is often done by web developers and website designer, that are highly qualified to code and manage the technical aspect of web management.

The responsibility of content is often overlooked by businesses. A meta description is a one-line summary of the webpage and therefore it plays a very important role in bringing visitor conversion.

Furthermore, the turnaround time for creating smaller descriptions is quite high. Typically, any website has around 20 links connecting to other web pages and each web page requires its unique meta descriptions.

Also as the businesses grow to expand, diversify, there are additions to the web links associated and hence bringing the need for updated meta descriptions. Meta descriptions are largely ignored because, despite holding the importance, it is quite an elaborate task, leading to much higher effort than the outcome.

Chapter 4: Objectives of the Study

The objective of this study is to address the major roadblocks to good website conversations, ensuring that businesses are able to generate appropriate and search-friendly descriptions with least effort and least turnaround time.

This study is aimed at building an approach to leveraging machine learning techniques to extract automatically extract meta-descriptions that are search-friendly and also within the permitted description limit and appropriately describe the webpage.

This will be achieved with the help of text summarization models in machine learning to reduce a corpus of text into the required length. By reducing the turnaround time and increasing the appropriateness, this approach can create a use case for the content writers and the web developers.

Chapter 5: Project Methodology

In this study, Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is adopted as the approach for building the machine-generated search-friendly meta description.

The CRISP-DM methodology is the process model consisting of six phases that describe the life cycle for any data science project. (Wirth, 2000) These six phases are:

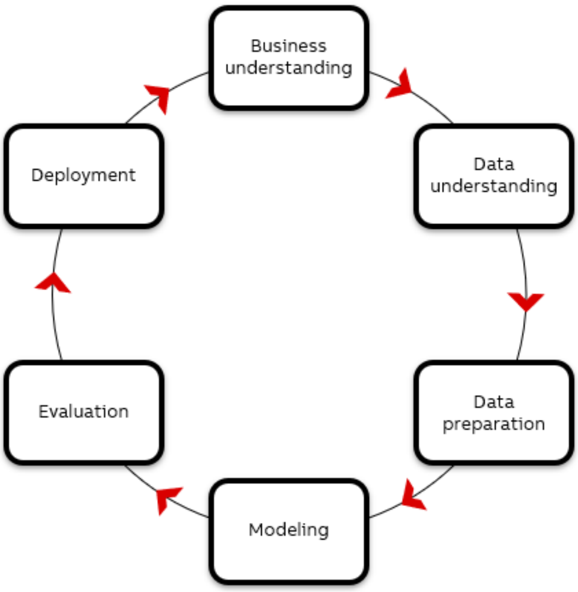


Figure no. 5

Business Understanding

The first step of Business Understanding phase helps focus on identifying the objectives of the project. By determining the business objective, it helps assess the situation, define goals for data mining, understand the project plan and nature of the business and domain.

Data Understanding

The second step in the modelling process is data understanding phase. At this stage, businesses share more information about the data. The aim is to identify, gather, and examine the data. This also includes exploring the data and understating various labels and features. It also included analyzing the data for potential quality issues which may later hamper the quality of the results. This includes looking at data types and sources and whether it is structured and unstructured data and plan for the approach.

Data Preparation

This phase is considered to be the longest and most crucial phase as it sets the course of the next steps and readies the data for the evaluations. The select data is cleaned and sanitized for any outliers, anomalies, biases, missing information and constructed based on the domain to be either standardized or creating newer features. This would also include combining various other data sources if needed and re-formatted to perform the next operation.

Modelling

The analytics and machine learning technique are applied at the modelling stage, where the data models are created based on data and the business goals. These models could be based on Regression models, Classification or Segmentation models. These models help generate more accurate plans and budgets. Splitting the data in training and test data based on approach, and check the model feasibility, with re-treating the data and altering it if needed.

Evaluation

At this stage, the strength of the model is put to test. Depending on the desired output and business goal, the results and findings of the model are examined. This also helps in preparing for the next steps in the approach. The results of the model are assessed against business goals.

Deployment

This is the final stage of the lifecycle, where on the successful completion of the evaluation stage, the business deploys the model into live projects to examine the results. This also helps in planning for the future project, looking at the business benefits as they meet the desired goals and the roadmap for the scope.

Chapter 6: Business Understanding

Data science and machine learning have found newfound interest among the learners. The increasing demand for analytics experts has led to increasing demand for the learners flocking to gain knowledge about business analytics.

In the last few years, many universities and websites have introduced data science learning courses in their programs for both beginners and professional. As a result, today there are thousands of e-learning courses and university programs offered in India. There is a growing competition among the institutes and colleges offering these courses to reach the prospect learner. The institutes and universities are aggressively working on search engine optimization techniques to grab the attention of these prospect learners. For a renowned university, it is important to be visible on the search results and ensure that the prospect visits the website and enrol for the program. The task of a digital marketing team is not only ensuring discoverability but conversion. A meta description plays an important role in ensuring conversion.

The requirement is to build an automated process to increase the appropriateness of the meta description while also reducing the turnaround time in creating the meta description. A good meta description has a few characteristics namely:

- It appropriately summarizes the webpage content, giving an accurate impression of what to expect.

- The length of the description should not exceed 156 characters or 20-30 words.

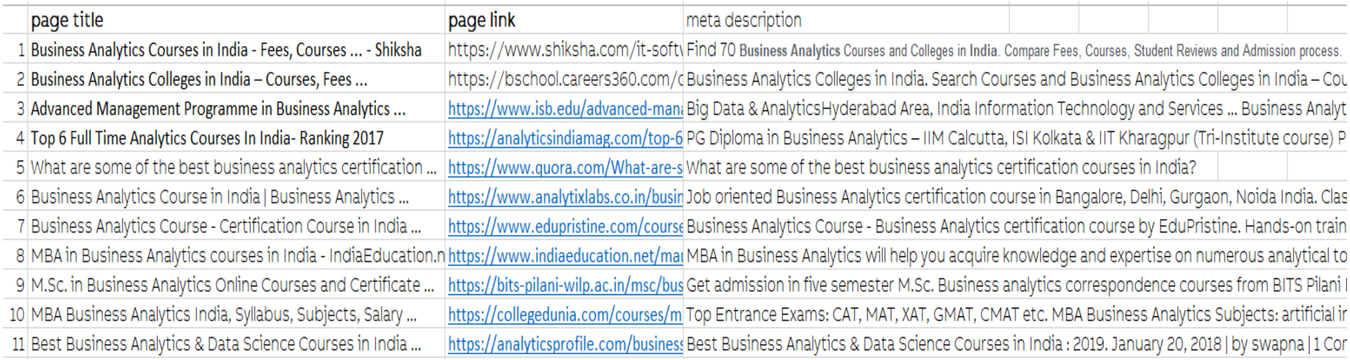
- The process of creating meta description should be automated to reduce the turnaround time and increase productivity

The description should include common and trending words used by other leading and search-friendly websites, to help the university build traction and attract the prospect learners to the website and convert more impressions.

Chapter 7: Data Understanding

To generate the desired description, we needed a corpus of words containing the desired keywords that are related to the common words and phrases used in search. The process starts with building a collection of sentences and descriptions. Also, these descriptions needed to be search-friendly. Since Google's search engine hold more than 90% of market share in internet search across devices such as desktop, mobile and tablets, we selected Google's search engine for generating the search results from the given search query.

Considering the business goal pertains to business analytics courses in India, we used 'Business' 'Analytics' and 'India' as the search terms for the search query on Google and pulled out the results of all 10 search pages.



	page title	page link	meta description
1	Business Analytics Courses in India - Fees, Courses ... - Shiksha	https://www.shiksha.com/it-softv	Find 70 **Business Analytics** Courses and Colleges in **India**. Compare Fees, Courses, Student Reviews and Admission process.
2	Business Analytics Colleges in India – Courses, Fees ...	https://bschool.careers360.com/c	Business Analytics Colleges in India. Search Courses and Business Analytics Colleges in India – Cou
3	Advanced Management Programme in Business Analytics ...	https://www.isb.edu/advanced-mana	Big Data & AnalyticsHyderabad Area, India Information Technology and Services ... Business Analyt
4	Top 6 Full Time Analytics Courses In India- Ranking 2017	https://analyticsindiamag.com/top-6	PG Diploma in Business Analytics – IIM Calcutta, ISI Kolkata & IIT Kharagpur (Tri-Institute course) P
5	What are some of the best business analytics certification ...	https://www.quora.com/What-are-s	What are some of the best business analytics certification courses in India?
6	Business Analytics Course in India | Business Analytics ...	https://www.analytixlabs.co.in/busin	Job oriented Business Analytics certification course in Bangalore, Delhi, Gurgaon, Noida India. Clas
7	Business Analytics Course - Certification Course in India ...	https://www.edupristine.com/course	Business Analytics Course - Business Analytics certification course by EduPristine. Hands-on train
8	MBA in Business Analytics courses in India - IndiaEducation.n	https://www.indiaeducation.net/ma	MBA in Business Analytics will help you acquire knowledge and expertise on numerous analytical to
9	M.Sc. in Business Analytics Online Courses and Certificate ...	https://bits-pilani-wilp.ac.in/msc/bus	Get admission in five semester M.Sc. Business analytics correspondence courses from BITS Pilani
10	MBA Business Analytics India, Syllabus, Subjects, Salary ...	https://collegedunia.com/courses/m	Top Entrance Exams: CAT, MAT, XAT, GMAT, CMAT etc. MBA Business Analytics Subjects: artificial ir
11	Best Business Analytics & Data Science Courses in India ...	https://analyticsprofile.com/business	Best Business Analytics & Data Science Courses in India : 2019. January 20, 2018 | by swapna | 1 Cor

Figure no. 7

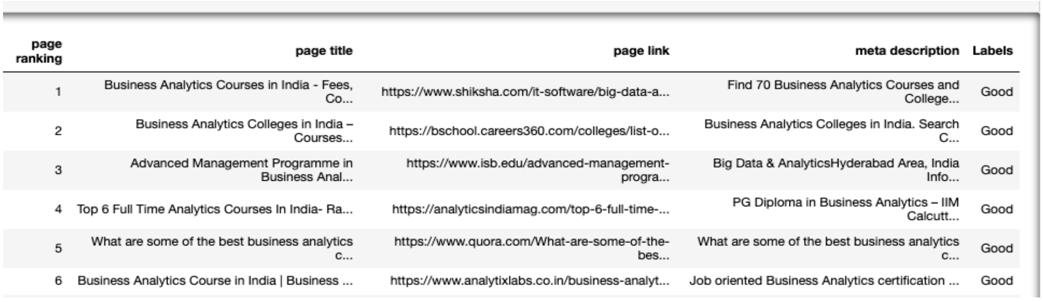
Our dataframe included top 100 observations that came as search result of the input words. It included the page title, the page link and their respective meta description. The aim is to build a corpus of words from the existing meta descriptions and summarize them together in a concise way to provide us the best keywords and the apt meta description.

Chapter 8: Data Preparation

Based on the business requirement and data understanding, we work on rule-based supervised learning methodology.

Various studies on page ranking and page views have shown that pages appearing on the first page or the top ten search results in Google search receive around 32% of the clicks of all the visitors. The click-through rates or the conversations drastically drop from the second page.(Brian Dean, 2019) One report further elaborated that 75% of the search clicks occur up to the third page, leaving very limited scope for the website falling after the third page or the 31st position. And it was also an interesting find that the visitors do not scroll beyond the fifth reach page or the 50th position.

This finding was basis of our data preparation part. Based on this industry-standard understanding, as mentioned above, since page visitors do not go beyond five search result pages , or the 50th search results, we created a new table marking the first 50 observations as Good and the remaining as Bad.



page ranking	page title	page link	meta description	Labels
1	Business Analytics Courses in India - Fees, Co...	https://www.shiksha.com/it-software/big-data-a...	Find 70 Business Analytics Courses and College...	Good
2	Business Analytics Colleges in India – Courses...	https://bschool.careers360.com/colleges/list-o...	Business Analytics Colleges in India. Search C...	Good
3	Advanced Management Programme in Business Anal...	https://www.isb.edu/advanced-management-progra...	Big Data & AnalyticsHyderabad Area, India Info...	Good
4	Top 6 Full Time Analytics Courses In India- Ra...	https://analyticsindiamag.com/top-6-full-time-...	PG Diploma in Business Analytics – IIM Calcutt...	Good
5	What are some of the best business analytics c...	https://www.quora.com/What-are-some-of-the-bes...	What are some of the best business analytics c...	Good
6	Business Analytics Course in India | Business ...	https://www.analytixlabs.co.in/business-analyt...	Job oriented Business Analytics certification ...	Good

Figure No. 8 (a)

The figure shows the top 6 observations with a new label based on their page ranking.

This rule of labelling of descriptions helped us distinguish between the meta description used by high ranking websites and the low-ranking websites. It also helped in filtering out the bottom 50 results, marked as ‘Bad’ in labels. The remining meta descriptions across the first 50 observations are to be concatenated together into a corpus of string.

The next step in data preparation was pulling out unwanted words and cleaning out the corpus to retain usable and required keywords. This process is called dropping stop words. Stop words are the commonly appearing words such as articles that do not carry any significance and because of their higher frequency in a paragraph, stop words often come as the highest frequency words.

In order to identify the right keyword and focus on these words, Stop words are often removed or dropped from the string corpus and the words appropriately reflect the word frequency.

The next step from stopword removal is getting the frequency of the words, to highlight the high-frequency words. It starts with building a dictionary from the frequency table using the text from the corpus and then splitting the corpus into sentences.

Creating good description required identifying the best descriptions, bringing them together as a corpus of sentences and finding the best way to summarize them. There are multiple approaches to summarization, but it starts with helping the machine identify the text, though tokenization and vectors.

The inbuilt method from Natural Language Toolkit (NLTK) was used. The NLTK tool works as the starting point of working on text. It helps break down the paragraphs in a corpus into small parts or chunks of words or phrases and sentences. Each part is called tokens and the process is called as tokenization. (Bird et al., 2009)

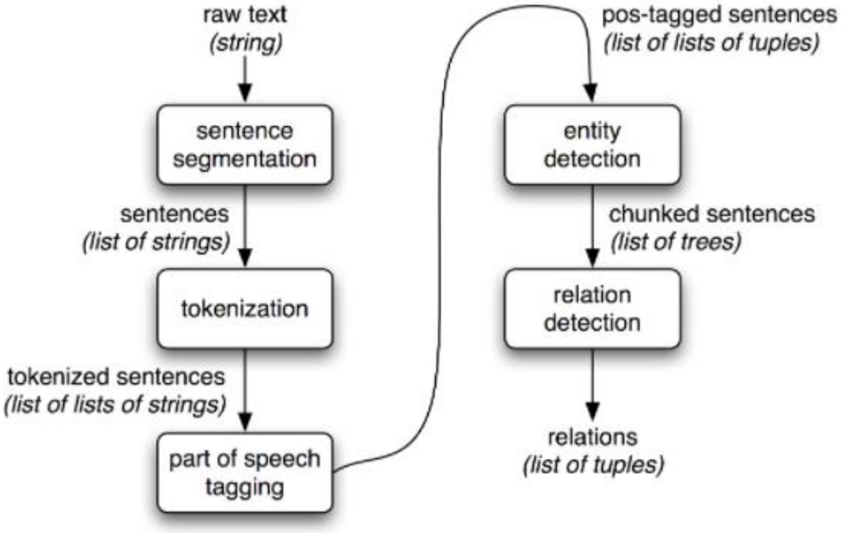


Figure 8 (b)

Chapter 9: Data Modeling

One of the common approaches to summarization is TF-IDF (term frequency–inverse document frequency). However, TF-IDF, was not beneficial in this approach. (Robertson, 2004)

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Figure 9 (a)

In the calculation of TF-IDF, the value of the words is indirectly proportional to the number of times it is repeated in the corpus. As a result, the higher the frequency of the words the higher changes of the words getting filters out in the resulting corpus. While the business requirement was to identify such high frequency words and define them as important keywords based on the frequency.

The requirement was to find the words with high frequency and identify them as important keywords for the description. We, therefore, deployed another approach of PageRank to score the tokenized sentences and started with building the term frequency for scoring the tokenized sentences.

To start with, we brought in the Term Frequency methodology and scored every sentence. We took the average score of these sentences to be the desired threshold. We picked the sentences for summarization based on the top five scored sentences.

The next step from here was summarizing and reducing the corpus of 50 meta descriptions into five sentences with the highest frequency scores based on the algorithm. We applied text summarization to summarize the pages.

There are two methodologies within the text summarization methodology namely the extractive model and the abstractive model. There has been academic work around the exercise especially Extractive methodology to condense sentences and phrased based on their importance. An extractive summarization technique consists of picking important lines or paras from the initial document and concatenating putting them in shorter form. It is calculated on statistical and linguistic features of sentences.(Allahyari et al., 2017) The purpose is to create a consistent and fluent abstract having only the most important points outlined in the record. Automatic text summarization is a familiar problem in natural language processing (NLP). Lately, there has been boom in the text data from a variation of sources. This publication is an extremely useful source of information, which needs to be constructively summarized to be practical. The main perspective to programmed text summarization are described.

For the purpose of text summarization, we used PageRank Algorithm. Page Rank is often described as an extractive summarization seeks to select a subset of the words or sentences in the existing document which best represents a summary of the document. PageRank a model for ranking sentences based on graph for text processing and is used in natural language processing applications. It uses two approaches that of "scoring" and the "recommendation".(Joshi, 2018) There has been a lot of research work on extracting the core of the text or summarizing text using PageRank. We run the PageRank algorithm to extract the top five sentences giving the highest scores.

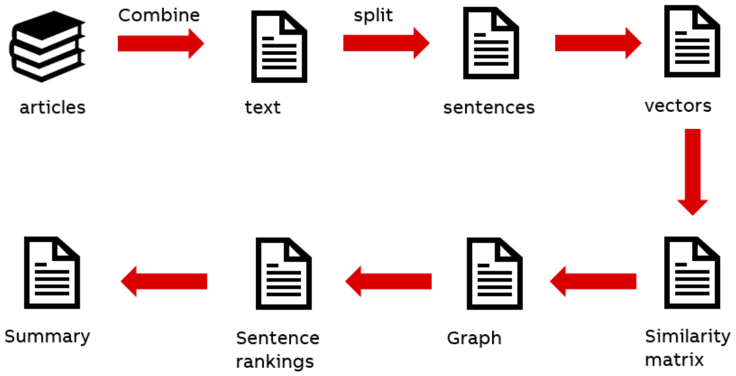
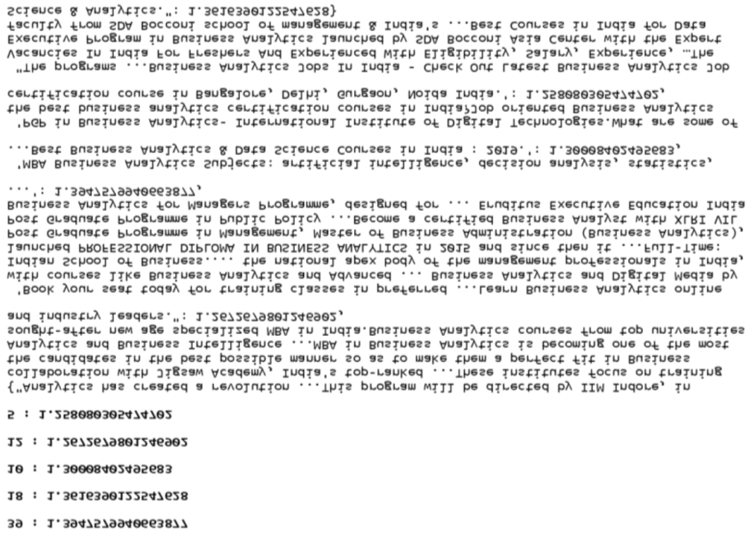


Figure no. 9 (b)

As shown in the image, the algorithm splits sentences and creates vectors based on context. Using Cosine index, it is possible to put them in a similarity matrix to find highly used sentences and word. As detailed above, the graph-based ranking provides a score to all the sentences, helping us find the top scored sentences.

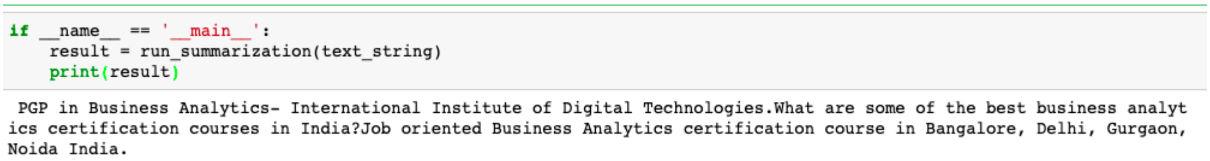


```
39 : 1.394757994066387

18 : 1.361639012254762

10 : 1.300084024956683

12 : 1.267267980124690

5 : 1.258080305474702

{"Analytics has created a revolution ...This program will be directed by IIM Indore, in
collaboration with Jigsaw Academy, India's top-ranked ...These institutes focus on training
the candidates in the best possible manner so as to make them a perfect fit in Business
Analytics and Business Intelligence ...MBA in Business Analytics is becoming one of the most
sought-after new age specialized MBA in India.Business Analytics courses from top universities
and industry leaders.": 1.2672679801246902,

 'Book your seat today for training classes in preferred ...Learn Business Analytics online
with courses like Business Analytics and Advanced ... Business Analytics and Digital Media by
Indian School of Business.... the national apex body of the management professionals in India,
launched PROFESSIONAL DIPLOMA IN BUSINESS ANALYTICS in 2015 and since then it ...Full-Time:
Post Graduate Programme in Management, Master of Business Administration (Business Analytics),
Post Graduate Programme in Public Policy ...Become a certified Business Analyst with XLRI VIL
Business Analytics for Managers Programme, designed for ... Eruditus Executive Education India
...': 1.394757994066387,

 'MBA Business Analytics Subjects: artificial intelligence, decision analysis, statistics,
...Best Business Analytics & Data Science Courses in India : 2019.': 1.300084024956683,

 'PGP in Business Analytics- International Institute of Digital Technologies.What are some of
the best business analytics certification courses in India?Job oriented Business Analytics
certification course in Bangalore, Delhi, Gurgaon, Noida India.': 1.258080305474702,

 "The programs ...Business Analytics Jobs In India - Check Out Latest Business Analytics Job
Vacancies In India For Freshers And Experienced With Eligibility, Salary, Experience, …The
Executive Program in Business Analytics launched by SDA Bocconi Asia Center with the Expert
faculty from SDA Bocconi school of management & India's ...Best Courses in India for Data
Science & Analytics.": 1.361639012254762}
```

Figure no. 9 (c)

Chapter 10: Data Evaluation

The PageRank provides us the ideal list of five high scores sentences as descriptions. The next the target is to achieve the best description in 150 characters of 15-25 words. We recreate a new corpus the using the high scored sentence and run another round of simple text summarization using the frequency methodology.



```
if __name__ == '__main__':
    result = run_summarization(text_string)
    print(result)
```

```
 PGP in Business Analytics- International Institute of Digital Technologies.What are some of the best business analyt
ics certification courses in India?Job oriented Business Analytics certification course in Bangalore, Delhi, Gurgaon,
Noida India.
```

Figure 10

Based on the modelling results the summarized and scored set of words are “PGP in Business Analytics Business Analytics Certification Course India Job Oriented Certification Course in Bangalore, Delhi, Gurgaon, Noida.”

The result is 18 words or 143 characters keywords, which could be strung together as a meaningful sentence making a search-friendly meta description, that includes the ideal set of keywords.

Chapter 11: Analysis and Results

The analysis of the approach shared a set of words for a meta description.

Bringing them all together, a good search friendly meta description within the required word limit could be “One of the best business analytics certification courses in India that offer job-oriented Business Analytics certification course in Bangalore.”

We believe using this approach to building machine-generated search-friendly meta descriptions for websites, a business could look at better click-through rates and conversions from the searchability.

Chapter 12: Conclusions and Recommendations for future work

This is a proof of concept for building a description by stringing the words together. It could be used to address major problem areas faced by web and digital teams in bringing appropriateness in the content by picking and deriving the most suited descriptions and keywords from search-friendly websites in similar space, offering similar product and service. It can also help is ensuring consistent and required quality of the description.

Automated machine learning-based approach should also help in reducing the turnaround time for creating meta descriptions by 5 times, from 8-16 hours to 1-2 hours only and eliminate the requirement for a specific domain expert just to build the content.

This could be found useful in building interdependence on other websites since the process starts from identifying the top-ranking pages and respective meta descriptions and summarizing them into the desired output. The next step to this study brings further independence in the task of description generation.

Chapter 13: Bibliography

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., D., E., B., J., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2017.081052

Bird, S., Klein, E., & Loper, E. (2009). Language Processing and Python. *Computing*.

Brian Dean. (2019). Here's What We Learned About Organic Click Through Rate. *Backlinko*. https://backlinko.com/google-ctr-stats

Craven, T. C. (2004). Variations in use of meta tag descriptions by Web pages in different languages. *Information Processing and Management*. https://doi.org/10.1016/S0306-4573(02)00121-8

Danny Sullivan, S. E. L. (2015). Meet RankBrain: The Artificial Intelligence That's Now Processing Google Search Results. *Search Engine Land*. https://searchengineland.com/meet-rankbrain-google-search-results-234386

Google: Search Quality Raters. (2020). *Search Engine Land*. https://searchengineland.com/library/google/google-search-quality-raters

Joshi, P. (2018). *An Introduction to Text Summarization using the TextRank Algorithm*. https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

Khan, M. N. A., & Mahmood, A. (2018). A distinctive approach to obtain higher page rank through search engine optimization. *Sadhana - Academy Proceedings in Engineering Sciences*. https://doi.org/10.1007/s12046-018-0812-3

Labra Gayo, J. E., de Pablos, P. O., & Cueva Lovelle, J. M. (2010). WESONet: Applying semantic web technologies and collaborative tagging to multimedia web information systems. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2009.10.004

Lee, Y., & Kozar, K. A. (2006). Investigating the effect of website quality on e-business success: An analytic hierarchy process (AHP) approach. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2005.11.005

Max Irwin. (2018). An Introduction to Search Quality. *OpenSource Connections*. https://opensourceconnections.com/blog/2018/11/19/an-introduction-to-search-quality/

Page, L., & Brin, S. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*. https://doi.org/10.1016/s0169-7552(98)00110-x

Raj Krishnan, S. T. (2007). *Improve snippets with a meta description makeover*. https://webmasters.googleblog.com/2007/09/improve-snippets-with-meta-description.html

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*. https://doi.org/10.1108/00220410410560582

Salton, G., Singhal, A., Buckley, C., & Mitra, M. (1996). Automatic text decomposition using text segments and text themes. *Proceedings of the ACM Conference on Hypertext*. https://doi.org/10.1145/234828.234834

Schröer, S., & Schröer, S. (2018). Search Engine Optimization (SEO). In *Quick Guide Online-Marketing für Einzelkämpfer und Kleinunternehmer*. https://doi.org/10.1007/978-3-658-15939-9_5

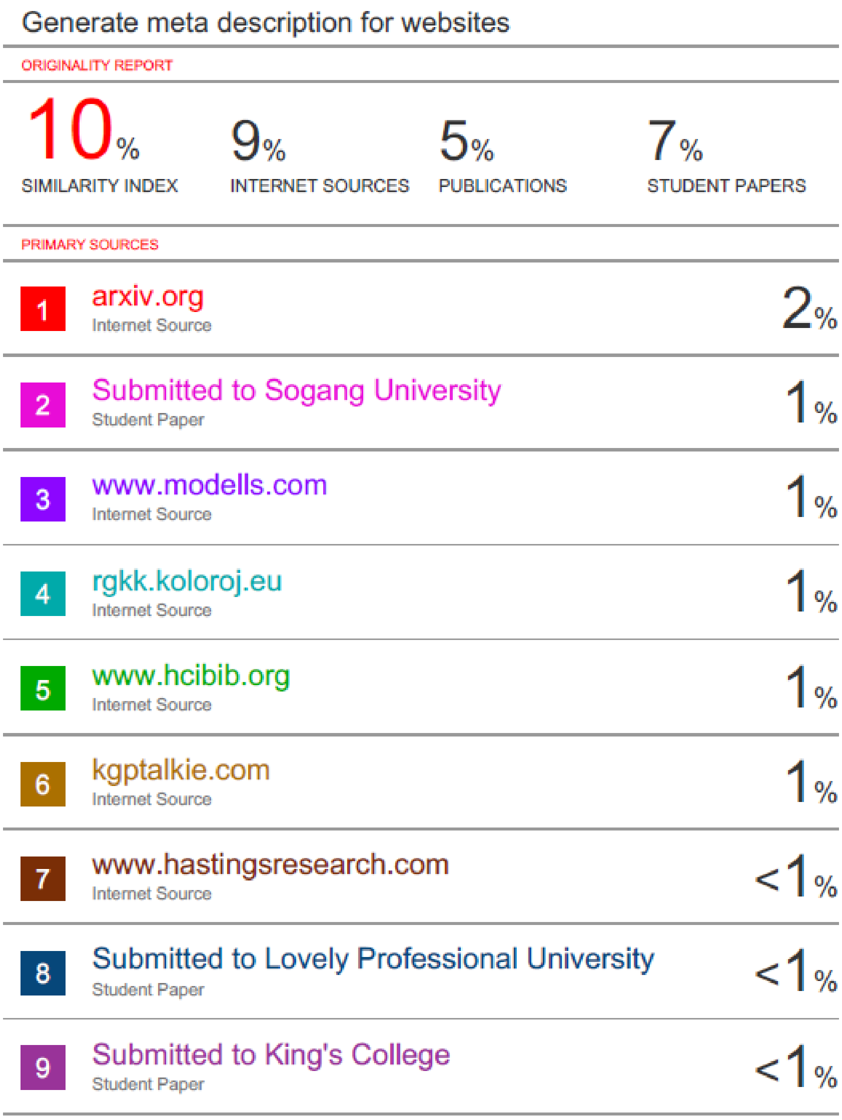
Shih, B. Y., Chen, C. Y., & Chen, Z. S. (2013). An empirical study of an internet marketing strategy for search engine optimization. *Human Factors and Ergonomics In Manufacturing*. https://doi.org/10.1002/hfm.20348

Swati, P. P., Pawar, B., & Ajay, S. P. (2013). Search Engine Optimization: A Study. *Isca.In*.

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*.

Appendix

Plagiarism Report[1]



Generate meta description for websites

ORIGINALITY REPORT

10%	9%	5%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	2%
2	Submitted to Sogang University Student Paper	1%
3	www.modells.com Internet Source	1%
4	rgkk.koloroj.eu Internet Source	1%
5	www.hcibib.org Internet Source	1%
6	kgptalkie.com Internet Source	1%
7	www.hastingsresearch.com Internet Source	<1%
8	Submitted to Lovely Professional University Student Paper	<1%
9	Submitted to King's College Student Paper	<1%

Publications in a Journal/Conference Presented/White Paper[2]

Any Additional Details

[1] Turnitn report to be attached from the University.

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.