**Group-1—P57**

**Task 1:- Research about “What is Topic Modelling”.**

**Topic Modelling**

1. Topic modelling, in the context of Natural Language Processing, is described as a method of uncovering hidden structure in a collection of texts.
2. Topic modelling is a form of unsupervised learning that identifies hidden relationships in data.
3. Topic modelling works in an exploratory manner, looking for the themes (or topics) that lie within a set of text data. There is no prior knowledge about the themes required in order for topic modelling to work.

There are two types of topic analysis techniques:

* **Topic modelling** is an ‘unsupervised’ machine learning technique, in other words, one that doesn’t require training.
* **Topic classification** is a ‘supervised’ machine learning technique, one that needs training before being able to automatically analyse texts.

In short, topic modelling algorithms churn out collections of expressions and words that it thinks are related, leaving you to figure out what these relations mean, while topic classification delivers neatly packaged topics, with labels such as Price, and Features, eliminating any guesswork.

* If you don’t have a lot of time to analyse texts, or you’re not looking for a fine-grained analysis and just want to figure out what topics a bunch of texts are talking about, you’ll probably be happy with a **topic modelling** algorithm.
* However, if you have a list of predefined topics for a set of texts and want to label them automatically without having to read each one, as well as gain accurate insights, you’re better off using a **topic classification** algorithm.

### Topic Modelling

Topic modelling refers to the process of dividing a corpus of documents in two:

1. A list of the topics covered by the documents in the corpus
2. Several sets of documents from the corpus grouped by the topics they cov

There are techniques for Topic Modelling are:

* Latent Semantic Analysis (LSA)
* Probabilistic Latent Semantic Analysis (pLSA)
* Latent Dirichlet Analysis (LDA)
* Non-Negative Metric Function
* Correlated Topic Model (CTM)

## **Latent Semantic Analysis (LSA)**

It is based on what is known as the [distributional hypothesis](https://en.wikipedia.org/wiki/Distributional_semantics) which states that the semantics of words can be grasped by looking at the contexts the words appear in. In other words, under this hypothesis, the semantics of two words will be similar if they tend to occur in similar contexts.

LSA computes how frequently words occur in the documents – and the whole corpus – and assumes that similar documents will contain approximately the same distribution of word frequencies for certain words. In this case, syntactic information (e.g. word order) and semantic information (e.g. the multiplicity of meanings of a given word) are ignored and each document is treated as a bag of words.

## **Latent Dirichlet Allocation (LDA)**

(Hint:-What is LDA means :==The word **‘Latent’** indicates that the model discovers the ‘yet-to-be-found’ or hidden topics from the documents. **‘Dirichlet’** indicates LDA’s assumption that the distribution of topics in a document and the distribution of words in topics are both Dirichlet distributions. ‘**Allocation’**indicates the distribution of topics in the document. )

* **Latent Dirichlet Allocation** (LDA) is used as a topic modelling technique that can classify text in a document to a particular topic. It uses Dirichlet distribution to find topics for each document model and words for each topic model.
* Latent Dirichlet allocation is a technique to map sentences to topics. LDA extracts certain sets of topic according to topic we fed to it. Before generating those topic there are numerous process that are carried out by LDA. Before applying that process we have certain amount of rules, facts that we considered.

Assumptions of LDA for Topic Modelling:

* Documents with similar topics use similar groups of words
* Latent topics can then be found by searching for groups of words that frequently occur together in documents across the corpus
* Documents are probability distributions over latent topics which signifies certain document will contain more words of a specific topic.
* Topics themselves are probability distribution over words

# **Limitations**

* Fixed K (the number of topics is fixed and must be known ahead of time)
* Uncorrelated topics (Dirichlet topic distribution cannot capture correlations)
* Non-hierarchical (in data-limited regimes hierarchical models allow sharing of data)
* Static (no evolution of topics over time)
* Bag of words (assumes words are exchangeable, sentence structure is not modelled)
* Unsupervised (sometimes weak supervision is desirable, e.g. in sentiment analysis)

**[What's the difference between Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA)?](https://www.quora.com/Whats-the-difference-between-Latent-Semantic-Indexing-LSI-and-Latent-Dirichlet-Allocation-LDA" \t "_blank)**

LSI (also known as Latent Semantic Analysis, LSA) learns latent topics by performing a matrix decomposition (SVD) on the term-document matrix.

LDA is a generative probabilistic model, that assumes a Dirichlet prior over the latent topics.

In practice, LSI is much faster to train than LDA, but has lower accuracy.

**Why we are using LDA for Topic Modelling?**

The only difference is that LDA adds a Dirichlet prior on top of the data generating process, meaning NMF qualitatively leads to worse mixtures. It fixes values for the probability vectors of the multinomials, whereas LDA allows the topics and words themselves to vary.

**Toolkits:**

With the development of topic models, several toolkits have appeared for the broad application of these topic models. The toolkits below are mainly used in natural language processing.

* **Gensim**
* **Scikit Learn**

**Creators**

**Anup Meshram**

**Daroga Ajahar Madar**

**Navnath Satre**

**Vaibhav Bansode**

**Sandeep Khandelwal**

**Task 2:- Research about web extraction of data on different social media Platforms**

## **What is web scraping?**

Web scraping is the process of collecting structured web data in an automated fashion. It’s also called web data extraction. Some of the [main use cases of web scraping](https://www.zyte.com/learn/what-is-web-scraping-used-for-learn/what-is-web-scraping-used-for/) include price monitoring, [price intelligence](https://www.zyte.com/learn/price-intelligence/), news monitoring, [lead generation](https://www.zyte.com/learn/lead-generation/), and [market research](https://www.zyte.com/learn/market-research/) among many others.

In general, web data extraction is used by people and businesses who want to make use of the vast amount of publicly available web data to make smarter decisions.

If you’ve ever copy and pasted information from a website, you’ve performed the same function as any web scraper, only on a microscopic, manual scale. Unlike the mundane, mind-numbing process of manually extracting data, web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet’s seemingly endless frontier.

## **The basics of web scraping**

It’s extremely simple, in truth, and works by way of two parts: a web crawler and a web scraper. The web crawler is the horse, and the scraper is the chariot. The crawler leads the scraper, as if by hand, through the internet, where it extracts the data requested. [Learn the difference between web crawling & web scraping](https://www.zyte.com/learn/difference-between-web-scraping-and-web-crawling/)and how they work.

### ****The crawler****

A web crawler, which we generally call a “spider,” is an artificial intelligence that browses the internet to index and searches for content by following links and exploring, like a person with too much time on their hands. In many projects, you first “crawl” the web or one specific website to discover URLs which then you pass on to your scraper.

### ****The scraper****

A web scraper is a specialized tool designed to accurately and quickly extract data from a web page. Web scrapers vary widely in design and complexity, depending on the project. An important part of every scraper is the data locators (or selectors) that are used to find the data that you want to extract from the HTML file - usually, XPath, CSS selectors, regex, or a combination of them is applied.

## **The web scraping process**

1. Identify the target website
2. Collect URLs of the pages where you want to extract data from
3. Make a request to these URLs to get the HTML of the page
4. Use locators to find the data in the HTML
5. Save the data in a JSON or CSV file or some other structured format

## **What is web scraping used for?**

### Price intelligence

In our experience, price intelligence is the biggest use case for web scraping. Extracting product and pricing information from e-commerce websites, then turning it into intelligence is an important part of modern e-commerce companies that want to make better pricing/marketing decisions based on data.

How web pricing data and price intelligence can be useful:

* Dynamic pricing
* Revenue optimization
* Competitor monitoring
* Product trend monitoring
* Brand and MAP compliance

### Market research

Market research is critical – and should be driven by the most accurate information available. High quality, high volume, and highly insightful web scraped data of every shape and size is fueling market analysis and business intelligence across the globe.

* Market trend analysis
* Market pricing
* Optimizing point of entry
* Research & development
* Competitor monitoring

### Alternative data for finance

Unearth alpha and radically create value with web data tailored specifically for investors. The decision-making process has never been as informed, nor data as insightful – and the world’s leading firms are increasingly consuming web scraped data, given its incredible strategic value.

* Extracting Insights from SEC Filings
* Estimating Company Fundamentals
* Public Sentiment Integrations
* News Monitoring

### Real estate

The[digital transformation of real estate](https://www.zyte.com/blog/web-scraping-real-estate-data-use-cases/) in the past twenty years threatens to disrupt traditional firms and create powerful new players in the industry. By incorporating web scraped product data into everyday business, agents and brokerages can protect against top-down online competition and make informed decisions within the market.

* Appraising Property Value
* Monitoring Vacancy Rates
* Estimating Rental Yields
* Understanding Market Direction

### News & content monitoring

Modern media can create outstanding value or an existential threat to your business - in a single news cycle. If you’re a company that depends on timely news analyses, or a company that frequently appears in the news, [web scraping news data](https://www.zyte.com/data-types/news-scraping-api/) is the ultimate solution for monitoring, aggregating, and parsing the most critical stories from your industry.

* Investment Decision Making
* Online Public Sentiment Analysis
* Competitor Monitoring
* Political Campaigns
* Sentiment Analysis

### Lead generation

Lead generation is a crucial marketing/sales activity for all businesses. In the 2020[Hubspot report,](https://www.hubspot.com/marketing-statistics?__hstc=234333761.f597576fe5cf26f374c4dcd373e79bb7.1578558839179.1607600764366.1607603907810.490&__hssc=234333761.1.1607603907810&__hsfp=480828235) 61% of inbound marketers said generating traffic and leads was their number 1 challenge. Fortunately, web data extraction can be used to get access to structured lead lists from the web.

### Brand monitoring

In today’s highly competitive market, it's a top priority to protect your online reputation. Whether you sell your products online and have a strict pricing policy that you need to enforce or just want to know how people perceive your products online,[brand monitoring with web scraping](https://www.zyte.com/brand-monitoring/) can give you this kind of information.

### Business automation

In some situations, it can be cumbersome to get access to your data. Maybe you have some data on your own website or on your partner’s website that you need in a structured way. But there’s no easy internal way to do it and it makes sense to create a scraper and simply grab that data. As opposed to trying to work your way through complicated internal systems.

### MAP monitoring

Minimum advertised price (MAP) monitoring is the standard practice to make sure a brand’s online prices are aligned with their pricing policy. With tons of resellers and distributors, it’s impossible to monitor the prices manually. That’s why web scraping comes in handy because you can keep an eye on your products’ prices without lifting a finger.

## **Library for Web Scraping**

## 1. Requests (HTTP for Humans)

## 2. lxml

## 3. Beautiful Soup

## 4. Selenium

## 5. Scrapy

6. MechanicalSoup.

7. Urllib.