Mathematics and Big Data

Text Mining - Part 1

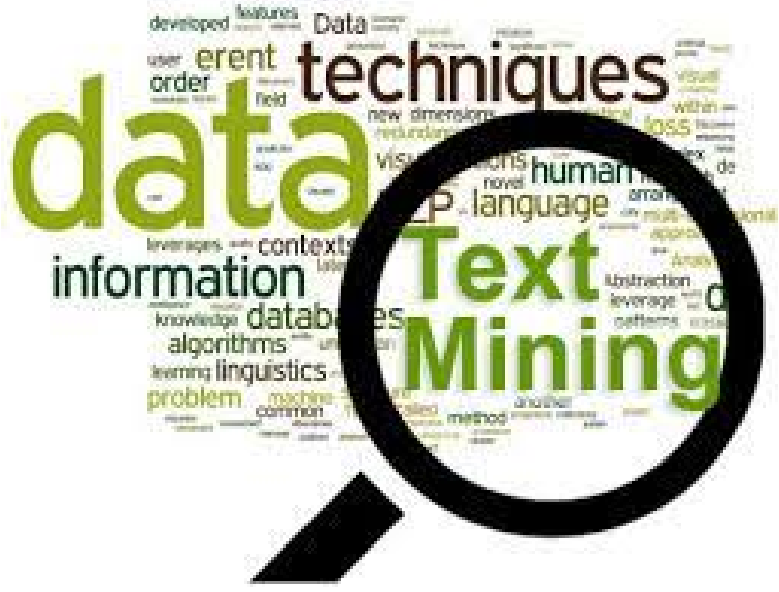
Sundus Zafar

Department of Mathematics
Autonomous University of Barcelona

1. Text Mining
   - Introduction
   - Why Text Mining ?
   - Applications
   - Challenges
   - Advanced projects
   - Text types
   - Definitions
   - Text preprocessing

Introduction

Text mining is a process of



- Automated extraction of interesting, non trivial information and knowledge from often large amount of unstructured text.
- Using computational methods and techniques to retrieve high quality information from text.

It is a subset of text analytics that is focused on applying data mining techniques in the domain of textual information using Natural Language Processing and Machine Learning.

Why Text Mining ?

Why do we need text mining?

- The information on the web is increasing rapidly. World's data doubles every 18 months.
- Demand of useful and reliable information from web in the shortest possible time by users.
- Due to vague specifications of required information by users the searching and extraction of information from the web texts using NLP technologies is required.
- An estimate of 90% world's data is in unstructured formats.

Applications

The vast applications of text mining can be viewed in the following categories:

- **Classification:** Documents, Stories, Books, Webpages etc.
- **Filters:** Spam, Emails, News, Adds etc.
- **Organization:** Search Engines, Repositories, Information Retrieval etc.
- **Knowledge:** Trends insight, Entity relations, Patterns, Predictions, Visualizations etc.
- **Information:** Entity associations, Recommender systems etc.

Challenges

Following challenges are commonly faced while working with text mining:

- Word and phrases ambiguity.
- Context sensitivity.
- High number of possible dimensions.
- Large textual database.
- Interpretation and comprehension of unstructured content.
- Cost of automated computational tools and methods.
- Statistical NLP: POS, Tokenization, Stemming etc.

Future Projects

With gained knowledge and some further knowledge in text mining following practical projects can be done (Potencial thesis topics):

- Spam email classifier/ filter system.
- News analysis.
- Twitter text analysis of trending topics.
- Sentiment Analysis of stock news.
- Stock trading recommender system.
- Recommender system of social networks.
- Content classification of official and social content.
- Fraud detection by pattern recognition.

Text types

- **Unstructured Text** It is categorized as **Qualitative data** and it cannot be processed or analyzed via conventional tools. It is present in several forms and in ever increainsg quantities such as:
  - Bussiness reports
  - Books
  - Various kinds of Bussiness documents
  - Digital libraries
  - Blog posts, articles
  - Wiki pages, messages, webpages
  - Social networking and social media sites
- **Structured Text** It is categorized as **Quantitative data**. It fits neatly within fixed fields e.g relational database, spreadsheets, dates, addresses, credit card numbers etc.

Structured vs Unstructured Text

Structured Data

What you find in a DB
(typically)

Unstructured Data

What you find in the 'wild'
(text, images, audio, video)

Definitions

Let's start with precise definitions of basic vocabulary used in text mining and NLP:

Definition

- A **Document** is the basic element. It is a unit of texual data e.g a sentence.
- A **Corpus** is a collection of documents. An unstructured set of texts.
- **Tokens** represent words.

Definitions

Definition

- **Terms** represent a single or multiwords units.
- A **N-grams** are set of co-ocurring words. They can be unigram, bigrams or trigrams etc.
- **Bag of words** consists of words from a sentence or document.

Definitions

Definition

- **Stopwords** are words to be excluded while text analysis.
- **Document representation** captures what document is about.
- **Document Term Frequency** numerically represents a document.
- **Document Term Matrix** numerically represents a collecton of documents.

Text preprocessing

In order to work with unstructured data it is important to do the required text preporcessing to remove the unnecessary text and make the necesary tansformations, e.g,

- Converting all letter to upper or lower case.
- Removing numbers or converting them.
- Removing punctuations, white spaces, stop words etc.
- Removing Nan values etc

Text Preprocessing with Python

```
>>> import re
>>> import string

>>> document = [___add document here with numbers, special
 characters and lower and upper case words___]

# Change the whole document to lower case
>>> doc_lower = document.lower()

# Change the whole document to lower case
>>> doc_upper = document.upper()

# Remove the numbers from the document
>>> doc_numbers = re.sub(r'\d+', '', document)
```

Text Preprocessing with Python

```
>>> import re
>>> import string

#Remove the set of symbols from the document
>>> doc_symbols = document.translate(string.maketrans("","")
string.punctuation)

>>> doc_symbol = re.sub('\[[^]]*\]', '', document)
```

Bag of Words with Python

```
>>>import sklearn
>>>import sklearn.feature_extraction
>>>from sklearn.feature_extraction.text import
   CountVectorizer

>>>corpus = [doc_1, doc_2, doc_3, ...]

>>>vec = CountVectorizer()
>>>X = vec.fit_transform(corpus)
>>>print(vectorizer.get_feature_names())
```