MES’S WADIA COLLEGE OF ENGINEERING, PUNE

Honors* in Artificial Intelligence and Machine Learning Fourth year of Engineering

410302: Machine learning Laboratory

NAME OF STUDENT:	**CLASS: BE**
SEMESTER/YEAR: VII	**ROLL NO:**
DATE OF PERFORMANCE:	**DATE OF SUBMISSION:**
EXAMINED BY: Dr. N. F. Shaikh	**EXPERIMENT NO: 04**

TITLE: Analysis on Twitter text data Perform text pre-processing, Apply Zips and heaps law, Identify topics

AIM/PROBLEM STATEMENT: Perform text pre-processing by applying zip and heap law

OBJECTIVES:

- To understand text pre-processing
- To understand the zip and heap law

OUTCOMES:

- Processing a Twitter Text
- Design zip and heap law

PRE-REQUISITES:

1. Theoretical knowledge of text pre-processing

2. Different pre-processing laws overview

THEORY:

Introduction

In recent days with the explosion of Big Data there is a large demand for organisations and data scientists to perform information extraction using non-traditional sources of data. Research has shown that nearly 80% of data exists as unstructured text data, hence text analytics is fundamental in order to analyse the wealth of information available on chat transcripts, social media posts, reviews, news feeds etc.

What is Text Analytics?

Text analytics is the processes of synthesising unstructured data to help discover patterns and enable decision making. Until recent years text analytics had to be performed the old fashioned way i.e. eyeballing and manual categorisation of text, which is inefficient and time consuming. Also this is not a practice solution when dealing with millions of documents such as twitter data. Twitter data (also known as tweets) is a rich source of information on a large set of topics. This data can be used to find trends related to a specific keyword, measure brand sentiment or gather feedback about new products and services. This post will provide a step by step guide for text analytics on twitter data.

How to perform text analytics on twitter data?

The steps involved in text analytics are:

Step 1: Collect tweets

Step 2: Pre-process tweets

Step 3: Apply sentiment analysis

Step 4: Apply named entity recognition

Step 5: Cluster tweets

Step 6: Visualise analysis

Step 2: Pre-Process tweets

Here we parse the response from the twitter API into a structured table. The response from twitter streaming API's is in the below format:

The detailed code for parsing the output from the twitter API is below. The output is structured into key fields such as "doc_id", "username", "text" etc. The parsed response can be stored in a database or a JSON file.

Zipf's Law

Zipf's Law is first presented by French stenographer Jean-Baptiste Estoup and later named after the American linguist George Kingsley Zipf. Zipf's Law states that a small number of words are used all the time, while the vast majority are used very rarely. There is nothing surprising about this, we know that we use some of the words very frequently, such as "the", "of", etc, and we rarely use the words like "aardvark" (aardvark is an animal species native to Africa). However, what's interesting is that "given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.

The **application of Zipf's Law** can be seen in most of the Natural Language Processing algorithm and in the Text Compression.

- In the NLP, Zipf's law is used to generate the AI powered text that looks like a natural text and contains the words that a normal human use.
- For compressing a text we cannot simply remove the words of our choice, so the Compression algorithm make use of Zipf's law that compress the most frequent word in the text and doesn't alter the words that occurs least frequently. We got the basic statistical approach of how Zipf's law is used.

Heaps' Law

The law can be described like as the number of words in a document increases, the rate of the count of distinct words available in the document slows down.

The documented definition of **Heaps’ law** (also called **Herdan's law**) says that the number of unique words in a text of n words is approximated by

$$V(n) = K n^\beta$$

where K is a positive constant and β is between 0 and 1. K is often upto 100 and β is often between 0.4 an 0.6.

The relation of Zipf's law with the Heaps' law is observed like as the length of the document(words in the document) keeps on increasing then after certain point we can see that no much unique words are added to the vocabulary.

Question:

1. What is Twitter Sentiment Analysis ?

2. Explain the different steps in data processing.