Predictive text mining application

Mohammad W. Ullah
Student No: 301369145
Manjur Rahaman
Student No: 301269270

Topic:

We would like to develop a text prediction algorithm similar to what we see in our mobiles.

Data Source:

For this project, the starting point is to get a big corpus. The corpus should contain proper sentences, for example, Shakespeare corpus or Reddit comments corpus or twitter or wiki data. To make variations in the sentence structure different types of data sources can be mixed together. We aim to experiment with many possible data sources to figure the best possible learning materials for our model.

Analysis:

The first step would be to clean and create word tokens (we learned in the first assignment). We will be using Markov Chain for storing the token probabilities and to forward from one step to the next.

Product:

An ideal product will be to create a web page where anyone can go and type any word or sentence to get a list (top 10) of next most probable words. It can also include a word cloud to represent the probability.

Technology:

We are going to use Pyspark for programming. It will greatly help at the tokenization stage to process the large volume of the text corpus. The project will help to learn about Natural Language Processing (NLP) techniques. We also wish to use Flask as our web front end.

Alternative Project Idea:

A music recommendation system based on the audio signal of a playlist. The goal is to extract audio features (signals) from the songs (mp3, wave etc) of an existing playlist of an user then compare that signal with a test playlist or songs and finally recommendation will be made based on the similarities found, if any. (Please note that we have this two ideas but we do not know how feasible the idea of music recommendation based on audio signal analysis is. We would appreciate your suggestion regarding the project choice.)

Conclusion:

This is a well-suited project for Big Data programming. The amount of text is abundant and any kind of structured sentence can be used for the project. It will be a good way to learn and explore big data tools. Data analysis will include data extraction, cleaning, mining, performance optimization, NLP, machine learning, visualization and finally web-app development.