Natural Language Processing

Content

- Collection of three main topics of high recent interest.
  - Search engines (Crawling, Indexing, Ranking)
    - Language Modelling
    - Text Indexing and Crawling
    - Relevance Ranking
  - Text Processing (NLP, NER, Sentiments)
    - Natural Language Processing
    - Named Entity Recognition
    - Sentiment Analysis
    - Summarization

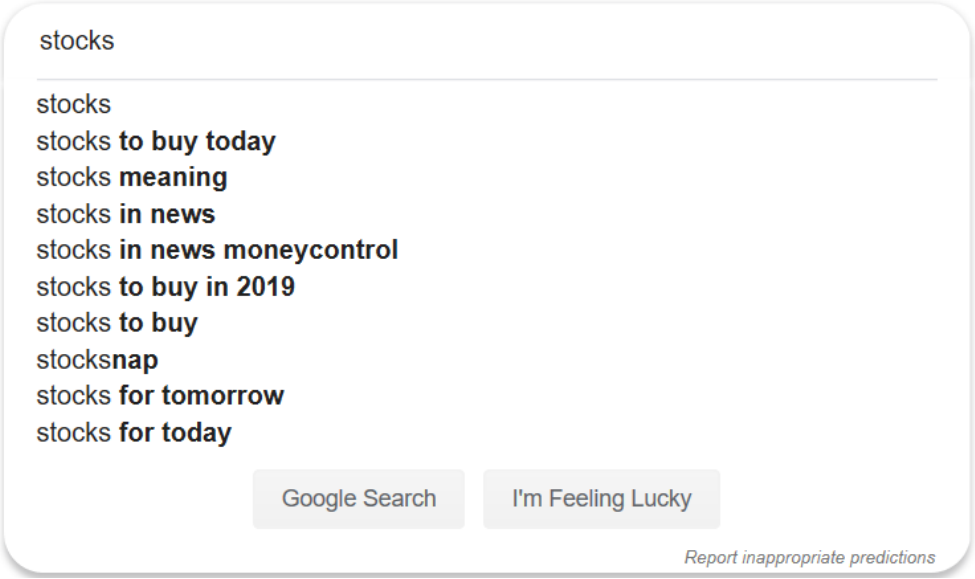
Relationship with Other Courses

- Fundamentals of Probability and Statistical Methods
  - These are the basics. You need them to understand the theoretical background for the course.
- Advanced Machine Learning, and Methods and Algorithms in Machine Learning
  - Lots of tasks in the course use ML algorithms and classifiers.
- Engineering Big Data with Python and Hadoop Ecosystem
  - The tasks that we will discuss, especially ones related to search engines can't be done on single machines.
  - We need Big data systems as the basic infrastructure for running the search engine algorithms.

Language Models

- Essential Probability and Statistics
- N-gram Models of Language

Next Word Prediction



How to do Word Prediction ?

- Use the previous N-1 words in a sequence to predict the next word
- Language Model (LM)
  - unigrams, bigrams, trigrams,...
- How do we train these models?
  - Very large corpora
  - Corpora are online collections of text and speech
    - Wall Street Journal
    - Newswire

N-grams

- Assume a language has V unique words in its lexicon, how likely is word x to follow word y?
- Simplest model of word probability: 1/V
  - Alternative 1: estimate likelihood of x occurring in new text based on its general frequency of occurrence estimated from a corpus (unigram probability)
  - popcorn is more likely to occur than unicorn
- Alternative 2: condition the likelihood of x occurring in the context of previous words (bigrams, trigrams,…)
  - mythical unicorn is more likely than mythical popcorn