N-Gram Model

Neelkumar S Bhuva
Yashwanth kumar Pamidimukkala

DOMAIN :

Natural Language Processing (NLP)

GOAL :

Predicting n^{th} word in a sentence of (n-1) words using probabilistic model. Ex : Turn in your homework … The 5^{th} word here can be by/tomorrow/on/today… Our goal here is to efficiently predict the next word based on the probability distribution learnt. Predict the word with the highest probability.

APPLICATIONS :

- Identify words in ambiguous input such as speech recognition or handwritten word recognition.
- Autofill sentence - Search engine, messaging application etc.
- Auto spell check. Ex: “I have a cat” is much more probable that “I have a caj” which is a spelling error.
- Machine Translation - Translating from any language to English.

DATASET :

- We have requested for the google dataset https://catalog.ldc.upenn.edu/LDC2006T13 for N-gram prediction.
- We also have Davies, Mark. (2011) N-grams data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.ngrams.info on November 11, 2017. The data here contains 1 word to 5 word sentences with frequency (count) of occurrence of each sentence.
- https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip. The data here contains sentences from news, blogs, and social media. We need to form N-grams using these files.

LITERATURE/BACKGROUND :

https://web.stanford.edu/~jurafsky/slp3/4.pdf

Task : Compute P(w|h), the probability of a word w given some history h.
One way to estimate this probability is from relative frequency counts: “Out of the times we saw the history h, how many times was it followed by the word w”. But, it turns out that even the web isn’t big enough to give us good estimates in most cases.

Intuition : The intuition of the N-gram model is that instead of computing the probability of a word given its entire history, we can approximate the history by just the last (n-1) few words.This is called Markov assumption.

SOFTWARE :

We plan to implement the N-Gram model using python language.
Libraries: Pandas, Numpy etc

CONTRIBUTION (TENTATIVE):

Data Acquisition and Cleaning: Neel
Data Preprocessing: Yashwanth
Modelling, Prediction and Testing: Together