Homework # 4

November 10, 2020

1 Bigram Language Model

Build a bigram language model on the data "J. K. Rowling - Harry Potter 1 - Sorcerer's Stone".

(1) Tokenize the data into a list of words using the function **nltk.wordpunct_tokenize**; convert all words to lower case; remove stop words and only keep alphabetic terms.

(2) Build a vocabulary from the tokens, how many unique words do you find?

(3) Build a frequency table where the rows represents the word w_{t-1} and the columns represent one word afterwards w_t. Count the occurrence of each word pair $C(w_{t-1}w_t)$

(4) How many times doe the following words occur: "harry", "stone", "hagrid", "feeling", "living" (hint: what if you sum all the numbers in a row?)

(5) Calculate the following conditional probabilities $p(potter|harry)$, $p(said|harry)$, $p(knows|everyone)$.

2 Word2vec: Medical Transcripts

Understand medical notes is a challenging NLP problem. Lots of good application can be made if a machine can read doctors' notes and interpret the underlying medical conditions and severity. In this exercise, you are presented a simple data of 5000 medical cases "**medicaltranscriptions.csv**". Each case has the transcript and the associated medical specialty. Please

(1) For the "description" of each individual, use "word_tokenize" function from nltk and convert the corpus into a list of words.

(2) Create a vocabulary containing all words appear in the descriptions. Count the number of total occurrence of each word. List the top 10 words that has the highest occurrence. Are those words related to medical terms?

(3) Convert the words in question (2) vocabulary to continuous vectors, using the pretrained word to vector dictionary "PubMed-and-PMC-w2v.bin". You may download the data from `http://evexdb.org/pmresources/vec-space-models`. Can all words find the corresponding vector representations?

(4) Calculate the cosine-similarity of the following word pairs: "allergy/allergic"; "heart/lung"; "water/-heart". Do the similarity measures make sense to you?