Yaffa Atkins
CUNY - Data Science
LING 83600/CSCI 76000
Prof. Levitan

HW1: Text Processing

The goal of this assignment is to give you hands-on practice writing code to process text. You will practice text segmentation and normalization in Python. You will also practice computing linguistic features. These skills are very useful for any computational linguistics tasks.

Deadline: Wednesday September 14, 11:59pm

Submission instructions: Submit a document with your answers to the HW questions (.pdf or .doc) as well as your code (.ipynb or .py). Your code must run with python3.

Here is a link to (optional) starter code for this HW:
https://colab.research.google.com/drive/1rjmnhUWfX4tGZu5Utxr56_F4F-A1wp3q?usp=sharing

Each student should make a copy of the notebook and of this document.

If you are not familiar with colab, here is a basic overview of its features:
https://colab.research.google.com/notebooks/basic_features_overview.ipynb

Download the data (you can use the starter code provided).
The 3 datasets you will need for this assignment are:

1. The complete works of William Shakespeare: shakespeare.txt
2. Excerpts from assorted news articles (headline + short summary): news.txt
3. Excerpts from the Switchboard corpus of phone conversation transcripts: swbd.txt

Word Tokenization
Tokenize each of the 3 datasets using the NLTK word tokenizer.
Count the number of tokens for each dataset, as well as the vocabulary size (number of unique tokens). Record the results in the table below. (1 pt)

Data	**Num Tokens (N)**	**Vocabulary Size (V)**	**Type to token ratio (TTR) V/N** (round to 3 decimal places)
Shakespeare	1229288	39531	0.032
News	33379	7247	0.217
SWBD	96221	5226	0.054

Print the top 10 most frequent tokens for each dataset. Record the results below. (1 pt)

Rank	Shakespeare		News		SWBD	
	token	freq	token	freq	token	freq
1	,	92610	the	1343	,	10323
2	.	81386	.	1201	:	5303
3	the	25536	to	855	.	5096
4	I	23305	,	673	I	3092
5	and	20145	of	658	the	1856
6	;	17357	and	637	and	1710
7	to	17285	a	636	that	1687
8	of	16983	in	608	you	1671
9	‘	16736	‘s	431	a	1568
10	a	13838	on	323	it	1501

Normalization

So far, we counted “unique” words without any preprocessing or normalization.
Word casing was not considered, so for example “The” and “the” are counted as distinct words.
Common words (like “a” or “the”) may not be as interesting as uncommon words. In NLP we call these words “stop words” and often remove them before processing text.

Punctuation marks are considered tokens, but we may want to filter them for some applications.

In this next step we will first normalize the tokenized text following these steps (in order):

1. Convert all text to lowercase
2. Remove punctuation (use string.punctuation for a list of punctuation marks)
3. Remove stopwords (use nltk.stopwords for English)

Note: order matters. Start with the tokenized text from the previous step and then apply these normalization steps in order.

Count the number of tokens and vocabulary size for the normalized text. Record the results in the table below. (1 pt)

Data	**Num Tokens (N)**	**Vocabulary Size (V)**	**Type to token ratio (TTR) V/N** (round to 3 decimal places)
Shakespeare	522361	32877	0.063
News	19466	6393	0.328
SWBD	38469	4724	0.123

Print the top 10 most frequent tokens for each dataset. Record the results below. (1 pt)

Rank	Shakespeare		News		SWBD	
	token	freq	token	freq	token	freq
1	thou	5865	's	431	‘s	1297
2	thy	4357	``	308	uh	1265
3	‘d	4305	''	260	know	1110
4	shall	3844	said	193	yeah	1030
5	‘s	3569	trump	141	uh-huh	785
6	thee	3374	new	90	n’t	784
7	lord	3162	would	75	–	540
8	king	3107	president	74	like	451
9	sir	2979	school	65	well	421

10	good	2950	n’t	65	think	419

Question 1: Having explored tokenization and normalization of these 3 datasets, what have you learned about the datasets? What other information would you like to learn about the data in the future? How do you think you might discover that information computationally? (1 pt)

Having explored the datasets using tokenization and normalization I have gained some insight into the data. It is interesting to note that the Shakespeare corpus is larger in both type and token size although it is not a surprise since the Shakespeare corpus contains the complete works of William Shakespeare, while the other two corpuses only contain experts. I do not find it surprising that Shakespear is the least varied since it is a compilation of stories written all by the same author and during the same time period. It is interesting to note that the news corpus is actually more varied than the conversions corpus since you would think news mainly discusses big stories in recent events, politics, the economy and the like while people discuss nearly everything and anything.

Before normalization the 10 most frequent tokens were all punctuation and stop words which is not surprising, all three of them had ‘.’, ‘,’ ‘the’, ‘and’, and ‘a’ in their top ten. After normalization I had a better understanding of what was contained within each dataset. Shakespeare is in old English so some words that may have been classified as stop words were not filtered out. I would like to know more than just the top ten most frequent words when looking at such a large database and specifically with Shakespeare I would include some old English in the list of stop words. The news dataset seems to be overwhelmingly about politics. I would be interested in the sentiment in the news articles and specifically in the titles compared to different politicians and political parties. To do this I would look for positive and negative words as well as the names of politicians and political parties and find the correlation. In the conversations dataset, I saw that many people use the words ‘like’, ‘uh’, ‘yeah’, and ‘uh-huh’ in their speech. This is not surprising as I see this happen in many of my day to day conversations as people think over what they are saying mid conversation and often do not articulate themselves on the spot as well as they would in writing. In speech people also like to talk about their opinions and what they think while also addressing the person they are talking with, possibly asking them their thoughts. Something that would be interesting to note are the main overlapping topics people discuss and their different sentiment towards them. I would do this by finding the most frequent keywords and look for the positive or negative words in close proximity to those keywords.

Feature extraction

We discussed Type-Token Ratio in class as a measure of complexity.

There are *many* other measures of text complexity and readability.
Here is a python library that implements several measures: https://pypi.org/project/textstat/
Compute the TTR and 3 additional features to study. (4 pt, 1 per feature)

For each measure (TTR + 3):

- Describe the measure
- Compute the feature for each of the 3 corpora and record the value in the table below
- Note: many measures are dependent on the corpus size. To ensure comparability across corpora, create data subsets of each corpus consisting of the first 20k tokens.

Data	**TTR**	**Flesch Reading Ease Score**	**Readability Consensus**	**McAlpine EFLAW Readability Score**
Shakespeare	0.0037	59.23	9th and 10th grade	62.9
News	0.00385	59.74	11th and 12th grade	26.7
SWBD	0.0035	80.62	5th and 6th grade	24.4

Question 2: What are the strengths and limitations of these measures? What have you learned about the corpora by extracting these features? (1 pt)

Flesch Reading Ease Score: This test measures how easy a text is to read and understand. It looks at the average number of words in a sentence and at the average number of syllables per word. A higher score means it is easier to read while lower means it is more difficult.

Strength: Number of syllables in a word and the length of a sentence certainly contributes to the readability of the text. Longer words and sentences tend to be more confusing.
Limitation: This test does not take into account words that are rare or words that are not are not pronounced the way they are spelled. These issues would cause a reader to have more difficulty with a text rather than the issues addressed.

I learned that the Shakespeare and News corpas have very similar scores which is interesting since I believe many people have a more difficult time reading Shakespeare since it is in old English and News is current and relevant so it should be less difficult to read. The SWBD corpa

has a significantly higher score which makes sense since people do not often use long words and sentences in conversation, rather they use simple words.

Readability Consensus: This test is based on a variety of different tests and calculates the estimated school grade level needed to understand the text.

Strength: A big strength of this test is it does not measure based off of one aspect, it uses multiple tests that each calculate the complexity of a text differently to calculate a final score. This means multiple factors are taken into consideration with this test.
Limitation: This measure does not focus on anything specific and when calculated, the score may not reflect which aspect of the text is easy or difficult to read.

What has this measure taught me about the corpa: It is interesting to note that this measure ranks the news corpa as two grades more complex than the shakespeare corpa. We see in practise that high school students are reading shakespeare in class and from my experience high school students are not reading the news very often.

McAlpine EFLAW Readability Score: This measure gives the text a score based on how easy a foreigner would be able to read it, the measure focuses on the number of miniwords and the length of the sentences.
Strength: One strength of this measure is foreigners will have a harder time following longer sentences so looking at the words in a sentence is helpful. Miniwords will be easier for foreigners to understand so taking those into consideration is a strength.
Limitation: Foreigners is a very general term, if the foreigners native language is more similar to English they will have a quicker time learning to understand English compared to those whose native language is quite different from English.
What has this measure taught me about the corpa: Shakespeare would be quite difficult for foreigners to read since it is in old english and many words there are not seen anywhere else so they most probably would never have encountered those words, in fact, it is also difficult for native english speakers to read. News will be easier for foreigners to understand since many people will know what is happening in the world so they will be familiar with the content of the news. It makes sense that SWBD would be quite easier for foreigners to understand since it is all conversational and therefore they have probably heard a lot of those words quite often.