NLP 220 Assignment 2

sttau

November 2023

1 Part 1

ISEAR corpus is an emotion classification given a text for joy, fear, anger, sadness, disgust, shame, and guilt.

The misspelled label ”guit” is converted to ”guilt.”

The tokenization for Spacy and NLTK are very similar, with small discrepancies between them. Changing the minimum token size for the vocab does not affect the graph since we ignore the infrequent tokens. The top 100 frequent token plot remains the same.

Emotion Name	**Max-length**	**Min-length**	**Avg-length**
Anger	101	2	24.546
Disgust	178	1	21.263
Fear	119	2	23.933
Guilt	159	1	24.012
Joy	122	1	19.581
Sadness	102	2	19.805
Shame	168	1	22.343

Table 1: Sentence Length Parameters by Emotion

Emotion	**NLTK Vocab Size Before**	**Spacy Vocab Size Before**
Joy	2346	2265
Fear	2937	2849
Anger	3125	3050
Sadness	2323	2241
Disgust	3155	3098
Shame	2664	2586
Guilt	2669	2589

Table 2: Vocabulary Sizes for Each Emotion

Emotion	Top Tokens
Joy	NLTK: friend (127), got (119), time (104), passed (101), felt (100) Spacy: friend (134), got (119), time (105), passed (101), felt (100)
Fear	NLTK: night (164), one (136), afraid (125), car (118), would (117) Spacy: night (166), afraid (125), car (118), fear (116), home (115)
Anger	NLTK: angry (189), friend (183), one (111), time (82), told (81) Spacy: angry (189), friend (188), time (82), told (81), got (79)
Sadness	NLTK: died (192), friend (175), sad (170), felt (131), time (98) Spacy: died (192), friend (180), sad (170), felt (131), time (99)
Disgust	NLTK: disgusted (149), saw (140), felt (112), people (109), one (104) Spacy: disgusted (149), saw (140), felt (112), people (110), friend (101)
Shame	NLTK: ashamed (189), felt (178), friend (128), one (101), time (87) Spacy: ashamed (189), felt (178), friend (133), time (89), told (77)
Guilt	NLTK: felt (216), guilty (186), friend (163), mother (118), one (101) Spacy: felt (216), guilty (187), friend (171), mother (120), time (88)

Table 3: NLTK and Spacy Top Tokens for Each Emotion

2 Part 2

Sentiment Analysis for tweets The values of the tweet are positive, negative, or neutral.

Sentiment	Percentage (%)
Neutral (0)	44.801600
Positive (1)	39.639567
Negative (-1)	15.558833

Table 4: Distribution of Sentiments

3 Part 3

A XML file of movie data Including title, year, rating, and description.// Outputting files to csv and Json

Table 5: Movie Details

Title	Year	Rating	Description
Indiana Jones: The Raiders of the Lost Ark	1981	PG	’Archaeologist and adventurer Indiana Jones is hired by the U.S. government to find the Ark of the Covenant before the Nazis.’
THE KARATE KID	1984	PG	None provided.
Back 2 the Future	1985	PG	Marty McFly
X-Men	2000	PG-13	Two mutants come to a private academy for their kind whose resident superhero team must oppose a terrorist organization with similar powers.
Batman Returns	1992	PG13	NA.
Reservoir Dogs	1992	R	WhAtEvER I Want!!!?!
ALIEN	1979	R	””””””””””””””””””
Ferris Bueller’s Day Off	1986	PG13	Funny movie about a funny guy
American Psycho	2000	Unrated	Psychopathic Bateman

4 Part 4

Using tree elements to find thriller movies and print Title and Year

Title	**Year**
ALIEN	1979
Ferris Bueller’s Day Off	1986
American Psycho	2000

Table 6: List of Movies

5 Part 5

Skipped for now for Assignment 3

6 Part 6

Comparing preprocessed tokenized books from NLTK libraries with their generic NLTK tokenizer tool with the raw text file.

Pretokenized NLTK: [’[’, ’Emma’, ’by’, ’Jane’, ’Austen’, ’1816’, ’]’, ’VOLUME’, ’I’, ’CHAPTER’, ’I’, ’Emma’, ’Woodhouse’, ’,’, ’handsome’, ’,’, ’clever’, ’,’, ’and’, ’rich’]

Tokenized NLTK: [’[’, ’Emma’, ’by’, ’Jane’, ’Austen’, ’1816’, ’]’, ’VOLUME’, ’I’, ’CHAPTER’, ’I’, ’Emma’, ’Woodhouse’, ’,’, ’handsome’, ’,’, ’clever’, ’,’, ’and’, ’rich’]

Number of different tokens: 1081
10 example token: carpet, wet., gentleman., brother-in-law, unknown., dances., wedding-day, alone., school-girl, Ceremonies

Not much difference between the preprocessed text and our self-processed. I observed small errors with our generic NLTK tokenizer missing hyphens or periods. If we eliminated a lot of infrequent terms we see less of a curve for zipf’s law and more of a linear line. Printing the POS we can still observe the law. The more we zoom in on the labels the more linear it is.

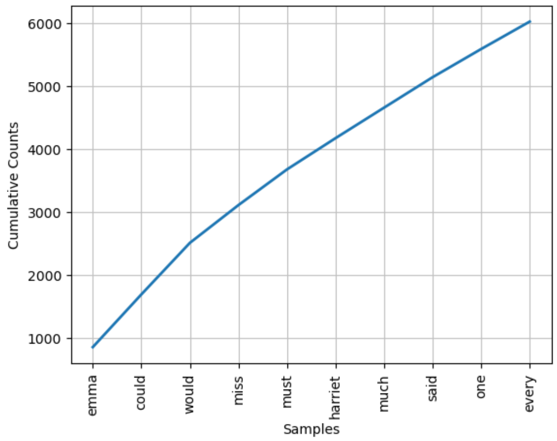

Figure 1: 10 Common Cumulative Token

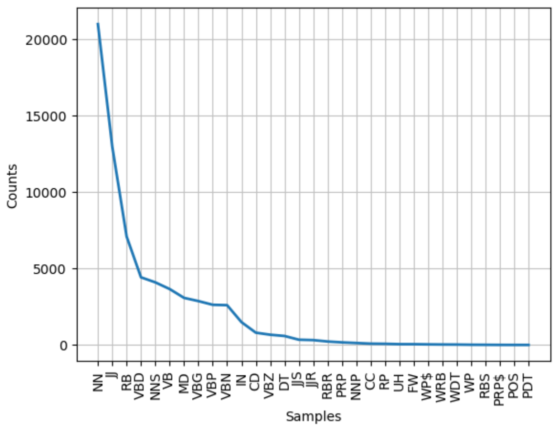

Figure 2: POS Frequency