Advaith Malladi

2021114005

Tokenization:

The process I followed:

- First, I replaced all the contractions in English in the corpus with their expanded form.
- I replaced all urls with URLHERE, mentions with MENTIONHERE, all hashes with HASHHERE, all numbers with NUMHERE, and all space patters with single spaces
- Then, I replaced all symbols except '!', '?', '.' with spaces.
- I replaced the above mentioned symbols with <eos> tags.
- Then, I replaced all address honorifics such mr, mrs, dr, in all their forms with ADDRESSHERE
- Then I replaced:
    - **Personal pronouns with 'personpronoun'**
    - **Reflexive pronouns with 'reflexpronoun'**
    - **Possessive pronouns with 'possespronoun'**
    - **Demonstrative pronouns with 'demonpronoun'**
    - **Interrogative pronouns with 'interpronoun'**
    - **Indefinite pronouns with 'indefpronoun'**
    - **Articles and with 'articlehere'**
    - Third person pronouns with 'thirdpronoun'
- Now, I split my entire corpus into a list of sentences where each sentence is a list of words

As we observe, the tokenization process I followed, especially replacing all pronouns with their categories helped me achieve great probability and perplexity in the statistical language modelling.

Observations in Perplexity:

Smoothing methods:

Pride and Prejudice corpus:

Knesser Ney smoothing: Perplexity Scores

Training: 211

Testing: 1056

Witten Bell smoothing: Perplexity Scores

Training: 893

Testing: 1345

Ullyses corpus:

Knesser Ney smoothing: Perplexity Scores

Training: 505

Testing: 2628

Witten Bell smoothing: Perplexity Scores

Training: 301

Testing: 1360

Neural Methods:

Pride and Prejudice corpus:

Training: 2290

Testing: 2343

Ullyses corpus:

Traning: 2587

Testing: 2740

ANALYSIS

- The best performance for Pride and Prejudice corpus was observed with Knesser Ney smoothing.
- The best performance for Ullyses corpus was observed with Witten Bell smoothing.
- **It does seem strange that the statistical model performed better. This can be attributed to the fact that I**

categorized all the pronouns in the statistical model, and I did not do so in the Neural Model.

- I did not categorize the pronouns in the Neural model because I used the Neural Model for generating outputs.
- If the pronouns are categorized, the Neural Model would not be able to generate outputs as it would not learn an pronoun, but would only learn pronoun categories