HW2

Note

- I really tried my best to submit a working model for 2.4, trying a couple dozen different sets of hyperparameters and methods but couldn't get anything to push past ~.45. Hopefully the extra credit fills in the 10 lost points for that?

Q1

★ 1.1.1 Use the helper function to create a frequency dictionary for the sentences in `en_corpus`. What are the 10 most frequent wordtypes?

- ('the', 34392), ('and', 24581), ('to', 15159), ('a', 10956), ('he', 9618), ('of', 9241), ('was', 7858), ('in', 6552), ('his', 5785), ('that', 5705), ('I', 5407)

★ 1.1.2 What are two limitations you observe in this simple tokenization method? Include examples that motivate these limitations.

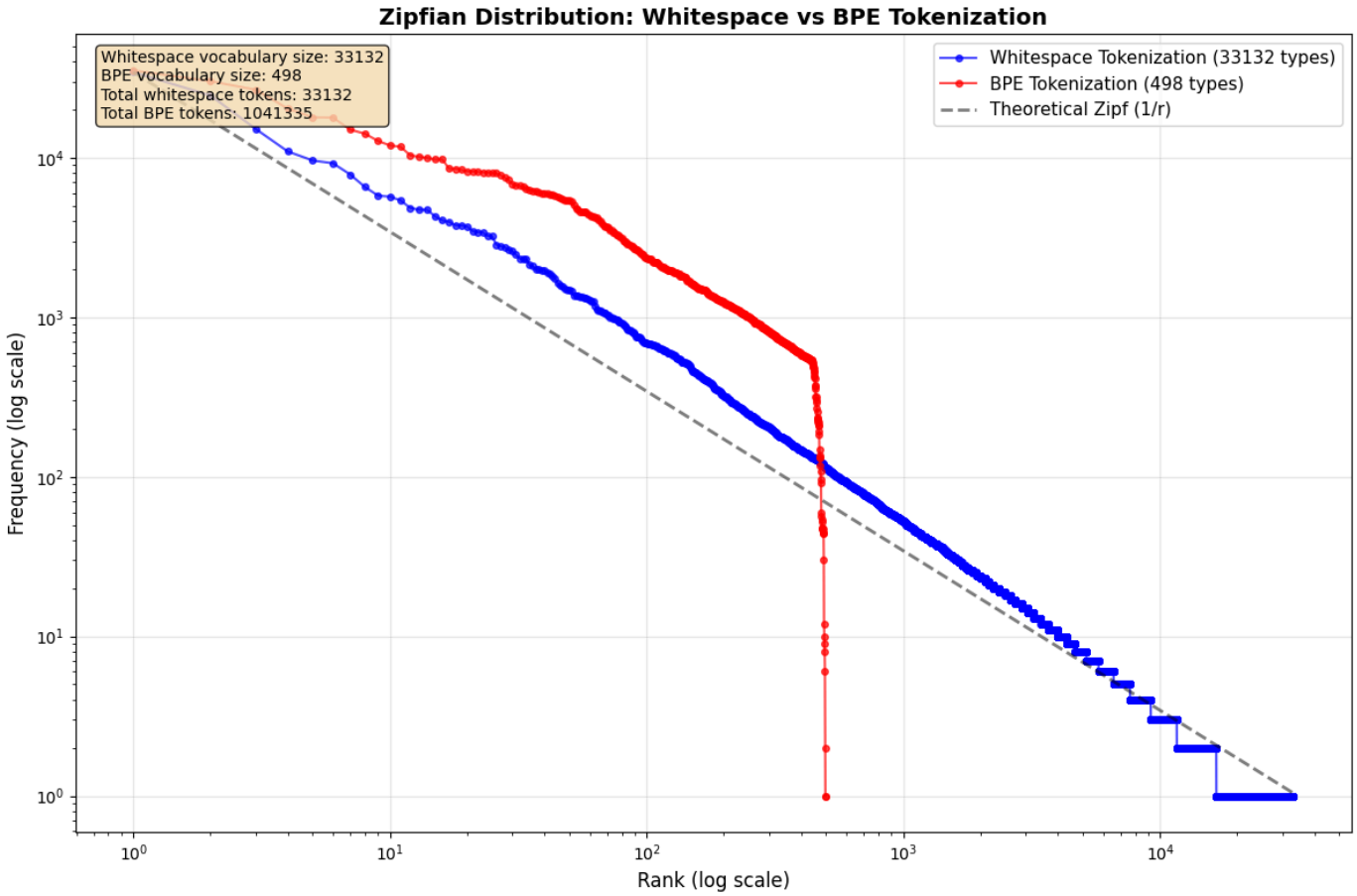
- This tokenization method is far too simple to capture a lot of useful information, especially with prepositions and verb phrases. This means meaning, relations, and more are all lost as it's simply counts.
    - As one example, "to" doesn't capture much information when separate from verb phrases: "come to", "belong to", ...
    - As another example, finding common encodings for terms such as the gerund "-ing" can allow the model to associate such a conjugation with active action.
- Another limitation with this tokenization method is that it results in a serious longtail. Uncommon words will appear very few times, making their sparse tokenization somewhat unuseful.
    - As one example, words such as "salacious" will have few tokens, making doing useful work with them hard (in addition to exploding the vocabulary).

★ 1.2.1 What are the 10 most frequent wordtypes from your BPE tokenizer?

- 'the': 34392
- 'and': 24581
- 'to': 15159
- 'a': 10956
- 'he': 9618
- 'of': 9241
- 'was': 7858
- 'in': 6552

- 'his': 5785
- 'that': 5705

★ 1.2.2 Generate a plot showing the Zipfian distribution of wordtypes (i.e., a log-log plot of frequency vs. rank). Compare the distributions from simple whitespace tokenization and BPE. Include the plots and your comparison in the report, make sure your plot is clearly labeled.



- BPE is far more efficient, capturing far more information and frequency with its most common tokens than that of the whitespace tokenization. In fact, while the whitespace tokenizer had over 33,000 tokens, the BPE was able to do it with only 500.

★ 1.2.3 Why might these distributions look different from one another?

- The whitespace approximately follows the "expected" 1/r relationship, while the learned rules are able to adapt to common affixes (-ing, etc.), morphological similarity (combined with the previous fact, walk, walk-ing, walk-ed, walk-s, walk-er, ... can all share a common root, which wasn't possible previously.), and breaking down of "longtail" uncommon terms into more digestible pieces.

Q2

★ 2.3.1 For each of the following context settings, list the highest-probability *negative* sample according to the priors you computed:

- Context as a bag of words.

    - `BOW | word = rowed | idx 6983 | p = 0.0010397244477644563`
- Context as neighboring tokens, $N = 1$
    - `N1 | word = pitied | idx 7835 | p = 0.000974905036855489`
- Context as neighboring tokens, $N = 2$.
    - `N2 | word = pinocchio | idx 1644 | p = 0.001003018580377102`

★ 2.4.1 Use `get_cosine_similarity` to explore similar wordtypes for different English words. Give three examples of pairs of similar wordtypes that make sense, and three that don't. What are possible reasons for a high similarity score between two seemingly unrelated words?

- Note that, for the above words, the G character is some sort of BOS/space tokenization.
- Three that make sense are...
    - "Stone" --> "Scold", referencing the more historical use of the word
    - "Kill" --> "shameful", alluding to immorality
    - "King" --> "Angel", although not super high on the list, may find some notion of "holiness" similar between the two
- Three that don't make sense are...
    - "Cat" --> "Fir", potentially having some cat/tree relation or, potentially, from typos of "fur"
    - "Man" --> "milk", while having the obvious "milkman" connection, seems far too strong (note: there seemed to have been a key error on "cow", which may explain this overconnection)
    - "kill" --> "ler", similar to above alluding to a "killer" connection of tokens, was surprising to me as this would result in "killler" spelled incorrectly(?)
- There are a few potential reasons behind seemingly unrelated words having high similarity scores.
    - Sparse training data may have uncommon words lying near each other.
    - Context provided in the training data may have a bias that doesn't encapsulate our traditional language.
    - Different window size might capture different amounts of context, which could seriously affect how words are defined as "related" (i.e. window=1 is going to only see adjacent preopositions, while window=10 will capture often unrelated phrases)

Q3

★ 3.1.1 Find three analogies in English that the model is able to solve correctly, and three analogies in English where the model gives a wrong answer.

- After throwing in some extraneous cases, we can see the following:
- Correct (relatively):
  - king:queen:: man:? -> she (expected: woman) -- gender is well captured across nouns
  - water:thirst::sleep:? -> o.O (expected: fatigue) -- somehow was able to categorize an emoticon (with *some* accuracy too)
  - finger: hand::toe:? -> shoe (expected: foot) -- part of body somewhat accurately captured

- Incorrect:
  - tiny : huge :: hot : big (expected: cold) -- the model can get stuck in loops involving original word context, despite the analogy equation being supposed to remove this. This implies that the learned model doesn't quite have enough parameters or training data to isolate "size", for example, as it's own portion of the vector.
  - tokyo: japan::london:? -> spain (expected: england) -- while locations are regularly grouped together, it doesn't quite have the resolution to distinguish between them in many cases
  - walk:walked::run:? -> had (expected: ran) -- past tense is strongly learned into the participle, likely overshadowing the actual past tense of the verb in question

★ 3.2.1 What happens when you translate each of these English words to Japanese/French, then back to English?

- FRENCH: penguin --> éléphant (elephant) --> elephant
- JAPANESE: penguin --> 海 (sea) --> ocean
-
- FRENCH: cheese --> fromage (cheese) --> sausage
- JAPANESE: cheese --> 美味しい (delicious) --> delicious
-
- FRENCH: sofa --> canapé (couch) --> sofa
- JAPANESE: sofa --> ソファー (sofa) --> bed
-
- FRENCH: jacket --> pantalon (pants) --> sleeves
- JAPANESE: jacket --> 上着 (jacket, coat) --> clothing
-
- FRENCH: website --> internet --> internet
- JAPANESE: website --> インターネット (internet) --> internet

Q4

★ 4.1.1 Report the 10 most frequent verb lemmas in the corpus.

1. say - 5771 occurrences
   2. go - 3952 occurrences
   3. have - 2978 occurrences
   4. come - 2910 occurrences
   5. see - 2285 occurrences
   6. take - 1984 occurrences
   7. do - 1676 occurrences
   8. give - 1510 occurrences
   9. make - 1279 occurrences
2. get - 1273 occurrences

★ 4.1.2 For the top 5 most frequent verb lemmas, show the distribution of their subject lemmas (i.e. by navigating the dependency tree from the root to find the `nsubj` item).

1. Verb: 'say' (5771 occurrences)
   Total subjects found: 4503
   Top 10 subjects:
   - he - 817 (18.1%)
   - she - 420 (9.3%)
   - they - 148 (3.3%)
   - man - 137 (3.0%)
   - King - 122 (2.7%)
   - woman - 110 (2.4%)
   - I - 86 (1.9%)
   - mother - 68 (1.5%)
   - father - 65 (1.4%)
   - son - 57 (1.3%)
2. Verb: 'go' (3952 occurrences)
   Total subjects found: 2641
   Top 10 subjects:
   - he - 596 (-
   - I - 299 (-
   - she - 255 (-
   - they - 247 (-
   - you - 112 (-
   - we - 90 (-
   - it - 68 (-
   - who - 52 (-
   - man - 41 (-
   - thou - 26 (1.0%)
3. Verb: 'have' (2978 occurrences)
   Total subjects found: 2489
   Top 10 subjects:
   - I - 468 (18.-
   - he - 447 (18.-
   - she - 221 (8.-
   - you - 190 (7.-
   - they - 165 (6.-
   - who - 158 (6.-
   - we - 82 (3.-
   - thou - 46 (1.-
   - one - 45 (1.-
   - it - 42 (1.-

4. Verb: 'come' (2910 occurrences)
   Total subjects found: 2334
   Top 10 subjects:
   - he - 308 (13-
   - they - 188 (8-
   - she - 123 (5-
   - I - 119 (5-
   - it - 98 (4-
   - you - 92 (3-
   - who - 56 (2-
   - man - 56 (2-
   - one - 43 (1-
   - time - 31 (1.3%)
5. Verb: 'see' (2285 occurrences)
   Total subjects found: 1493
   Top 10 subjects:

- he - 386 (25.9%)
- I - 235 (15.7%)
- she - 181 (12.1%)
- you - 150 (10.0%)
- they - 110 (7.4%)
- one - 44 (2.9%)
- we - 37 (2.5%)
- thou - 31 (2.1%)
- King - 25 (1.7%)
- who - 20 (1.3%)

★ 4.1.3 Based on your findings, do you observe any relationship between the types of verbs and the types of subjects they take (e.g., are certain verbs more likely to have animate subjects like 'man' or 'I', versus inanimate subjects like 'door' or 'it')? Discuss one interesting example.

- Some verbs are definitely dominated more than others by animate subjects. After taking a closer look at the top 5 lemmas (from the above part), we can see two drastically different examples:
  - "come" has 12.4% inanimate subjects, made up of tokens such as "it", "time", etc.
  - "see" on the other hand has far fewer, with none of the top 10 subjects being inanimate.
- This makes sense, marking a separation between more "active" verbs that involve a sentient subject, compared to more "passive" verbs that simply represent a transformation, such as "come".

Q5

ANSWERS

- A 4
- B 3
- C 7
- D 6
- E 8
- F 1
- G 2
- H 5

Notes:

- Object/Subject order reversed for pronouns, but sustained for nouns
- _/yau is sing/plurality for object
- ai/ge is sing/plurality for subject

Process (because below is so ugly):

- I first tried to fully understand the English options. As each sentence is a posessor(s)-posession(s) relationship, I immediately counted the total number of singular vs plural subjects and objects, and noted repetition of the nouns themselves.
- I then compared this to the counts of pronoun pairs in the mystery language. While I started this theory by seeing a small word (he/heau) in a similar location as "the", the 4/4 split ended up matching... the distribution of singular/plural OBJECTS. This then sparked the idea that there could be two pronouns, which matched up to the distribution of shai/shege.
- Finally, when tackling the nouns themselves, we have an easy "in" of the singular subject, plural object sentence (C 7). This made me realize that the suffixes somewhat match with the subjects, objects, and their respective pronouns, although by solving the roots for "manager" and "nurse" it was simple enough to begin to find the other uses of those words and classify sentences as such.

---

4 he (sing OBJ) (B C D H) (1 4 7 8)

4 heau (pl OBJ) (A E F H) (2 3 5 6)

5 shege (plural SUBJ)

3 shai (sing SUBJ)

A: p s ptauege yaijeech = schools arch

B: p p yaieege stauyau = archs teachers

C: s p ugheedai hyoogeshyau = manager's nurses

D: p p eegeege oogheedyau = managers tools

E: s s stauai eegeech = prof tool

F: s s yausai stauich = cousin teacher

G: p s hyoogesheege yauseech = nurses cousin

H: p p staueege ptauyau = teachers schools

EN STUFF:

1: singular singular (cousin teacher)

2: plural singular (nurses cousin)

3: plural plural (archs teachers)

4: plural singular (schools arch)

5: plural plural (teachers schools)

6: plural plural (managers tools)

7: singular plural * (manager nurses)

8: singular singular (prof tool)

ss: 2

sp: 1

ps: 2

pp: 3

cousin: 2 (s s)

teachers: 3 (+prof) (p(s) sp)

nurse: 2 (p p)

arch: 2 (p s)

school: 2 (p p)

manager: 2 (sp)

tool: 2 (ps)

prof: 1 (s)

if 4/4 split, ==> he/heau splits on the OBJECT'S plurality!