University of Pittsburgh

School of Computing and Information

CS2710

Natural Language Processing

Naive Bayes Classifier

Date: September 16, 2023

Jacob Hoffman

Advisors:

Michael Yoder, PhD

Part 1: results

```
The 10 most similar words to "juliet" using cosine-similarity on term-document frequency matrix are:
1: juliet; 1
2: romeo; 0.9899494936611666
3: capulet; 0.9899494936611665
4: pump; 0.9899494936611665
5: laura; 0.9899494936611665
6: pitcher; 0.9899494936611665
7: behoveful; 0.9899494936611665
8: hurdle; 0.9899494936611665
9: capulets; 0.9899494936611665
10: petrucio; 0.9899494936611665

The 10 most similar words to "juliet" using cosine-similarity on term-context frequency matrix are:
1: juliet; 1
2: lucius; 0.7547955309412351
3: warwick; 0.7372198370417355
4: servants; 0.7349817890175687
5: claudio; 0.7281556886727208
6: brutus; 0.7228154910841416
7: antonio; 0.7186173264439198
8: gloucester; 0.7183421383040222
9: fair; 0.7123146807943355
10: her; 0.7116414075333666

The 10 most similar words to "juliet" using cosine-similarity on tf-idf matrix are:
1: juliet; 1
2: pump; 0.9899494936611667
3: laura; 0.9899494936611667
4: pitcher; 0.9899494936611667
5: behoveful; 0.9899494936611667
6: hurdle; 0.9899494936611667
7: petrucio; 0.9899494936611667
8: heartless; 0.9899494936611667
9: searchers; 0.9899494936611667
10: tiberio; 0.9899494936611667

The 10 most similar words to "juliet" using cosine-similarity on PPMI matrix are:
1: juliet; 1
2: hist; 0.25901460016521716
3: capulet; 0.21865536712091427
4: romeo; 0.2051279969829155
5: lucio; 0.1765960610671714
6: nurse; 0.16470453828591025
7: tybalt; 0.164487063037831
8: booted; 0.1606389420523292
9: barnardine; 0.15990882858428623
10: laurence; 0.13555291409360448
```

```
The 10 most similar words to "king" using cosine-similarity on term-document frequency matrix are:
1: king; 1
2: sovereign; 0.8723889898823937
3: title; 0.8722882623156409
4: seat; 0.8624239750266073
5: subject; 0.86240703404077
6: kingdom; 0.8589420110465151
7: royal; 0.8581872075414547
8: lords; 0.8475107931059153
9: london; 0.8386619739210417
10: liege; 0.8382979584256983

The 10 most similar words to "king" using cosine-similarity on term-context frequency matrix are:
1: king; 1
2: people; 0.9423504308166016
3: prince; 0.9410910150632111
4: devil; 0.9317223008225002
5: french; 0.9316692275318647
6: ground; 0.9287778473508445
7: queen; 0.9280297646731829
8: sun; 0.9270199349274985
9: world; 0.9217903190507903
10: field; 0.9203647039826796

The 10 most similar words to "king" using cosine-similarity on tf-idf matrix are:
1: king; 1
2: sovereign; 0.8723889898823937
3: title; 0.8722882623156407
4: seat; 0.8624239750266072
5: subject; 0.8624070340407701
6: kingdom; 0.8589420110465155
7: royal; 0.8581872075414549
8: lords; 0.8475107931059154
9: london; 0.8386619739210417
10: liege; 0.8382979584256981

The 10 most similar words to "king" using cosine-similarity on PPMI matrix are:
1: king; 1
2: henry; 0.16306021058155107
3: enter; 0.14729434422168475
4: richard; 0.14290403369059768
5: queen; 0.13876735320211364
6: the; 0.13642367949442802
7: edward; 0.1294088038352298
8: lord; 0.12739324556313136
9: of; 0.12210306298940721
10: and; 0.12154713349421442
```

```
The 10 most similar words to "harry" using cosine-similarity on term-document frequency matrix are:
1: harry; 1
2: hoofs; 0.9817553661853936
3: scot; 0.9354385548260626
4: exchequer; 0.933799055647682
5: scots; 0.9334410060637159
6: westmoreland; 0.9322650589605471
7: fewer; 0.9179421297789901
8: horsemanship; 0.9179421297789901
9: eleventh; 0.9179421297789901
10: tun; 0.9179421297789901

The 10 most similar words to "harry" using cosine-similarity on term-context frequency matrix are:
1: harry; 1
2: gold; 0.7611892671073215
3: the; 0.7609223187940073
4: death; 0.751412327404711
5: blood; 0.7294797752534947
6: virtue; 0.7282883465284984
7: england; 0.7263923538261989
8: fair; 0.7257610511222345
9: youth; 0.7225881645001734
10: joy; 0.7218409261851143

The 10 most similar words to "harry" using cosine-similarity on tf-idf matrix are:
1: harry; 1
2: hoofs; 0.9817553661853937
3: scot; 0.9354385548260625
4: exchequer; 0.9337990556476818
5: scots; 0.9334410060637159
6: westmoreland; 0.9322650589605472
7: fewer; 0.91794212977899
8: horsemanship; 0.91794212977899
9: eleventh; 0.91794212977899
10: tun; 0.91794212977899

The 10 most similar words to "harry" using cosine-similarity on PPMI matrix are:
1: harry; 1
2: roy; 0.18774431317312112
3: cornish; 0.15401649797245487
4: percy; 0.15099890872792354
5: cousin; 0.14961951049315414
6: keepest; 0.14107099298350356
7: whencesoever; 0.14099777834455418
8: fitzwater; 0.1349441746363721
9: baron; 0.13144552682414545
10: appeals; 0.12976295528699122
```

1.3 - For the report:

- **In our term-document matrix, the rows are word vectors of D dimensions. Do you think that's enough to represent the meaning of words?**
  - Not really, because each row in a term-document matrix is giving a frequency of the specific word in each document within D, so this isn't necessarily providing any meaning about all of the words. Instead, it is helping us understand the relationship between each word and the documents, so for example this can assist us in ranking web pages (documents) during a search result.
- **Which vector space (term-document or term-context) produces similar words that make more sense than others and why do you think that is the case?**
  - Term-context seems to produce more similar words, and I think this is because the term-document matrix counts the frequency of the word in each document, but the term-context is comparing to a context window for each word, so the similarity is based on this range of words around the target/context words.
- **Consider any decisions you made in the prior sections when implementing your functions, such as whether you allowed a target word to co-occur with itself as a context word, and which window size you chose for the term-context matrix. How might any decisions you make impact our results now?**
  - Instead of scanning by line or with one large token stream for all documents, I chose to generate a token stream per document for the term-context matrices. I felt this would be more accurate overall because the aforementioned approaches would cause some invalid 'contextual positioning' (not sure what to call this) if the context window was cut short by a line or spanned across words that were only positioned near each other because of two document token streams being appended together.
  - Increasing the context window size can increase the context for each target word, which may provide further insight on the similarity between the target word and certain context words.
  - I allowed the target word to co-exist with itself as a context word, but I don't really think this is necessary. I'm not sure of a benefit that is gained from this, and it is always the top 1 similar word using cosine similarity metric, but this isn't very useful!
  - If I could make some changes, I would investigate making the matrices sparse, which would help optimize the performance of the program.

1.4 - *For the report:*

- **Redo the analysis in section 1.3 with ranked words and compare using tf-idf and PPMI with unweighted term-document and term-context matrices.**
- **Discuss findings from comparing approaches. Do some approaches appear to work better than others, i.e produce better synonyms? Do any interesting patterns emerge?**
  - I think that the term-context matrix with PPMI seems to produce the best synonym results, but both weighted options seem pretty good to me. Something that was a bit interesting is that tf-idf and PPMI don't seem to produce the same results (i.e. the

same words in the top 10 similar words).

- **How does weighting with tf-idf compare to using the unweighted term-document matrix? How does weighting with PPMI compare with using the unweighted term-context matrix? How does term-context/PPMI compare to term-document/TF-IDF? Include results and discussion.**
  - Unweighted term-document vs weighted tf-idf:
    - Raw frequency with a term-document matrix is not good for association between words because it is very skewed and not very discriminative. For example, the words “the” and “it” are not very informative about any particular word. Tf-idf helps to mitigate this.
  - Unweighted term-context vs PPMI:
    - When using PPMI, the association between two words is weighted by asking how much more the two words co-occur in the corpus than expected to appear by chance. PPMI helps to remove over-saturated words (that are found around all words), which can provide better similarity results.
  - term-context/PPMI vs term-document/tf-idf:
    - They are better for certain requirements. As mentioned before, term-document/tf-idf are good for rating the importance of a word in a document. term-context/PPMI are good for rating the importance of a word in the context of other words.

Part 2: results

For the report:

- **With that PPMI-weighted term-context matrix, find the vectors for identity labels in the provided list. Look at the top associated words (by PPMI) for at least 4 identity labels of your choice. Do you see any that may reflect social stereotypes? It is helpful to compare the top PMI words for certain identity terms with other related ones (such as men compared with women). Discuss and provide selected results in the report.**
  - I believe that it can be helpful to compare related identity terms, especially if you are researching something like social biases. This allows you to make a direct evaluation of the nth context-word order between two related groups, such as the given example (men and women). See results below.
- **Qualitative analysis: For at least 4 different identity labels, dig into the contexts that leads to high PMI association with other words, especially for any words that show social bias if you found that. 1st-order similarity: find specific examples from the dataset where an identity label occurs with a top-associated term that shows some social bias or does not. This might not occur; if not, you can look at 2nd-order similarity in which the two words occur with similar context words. Sample the contexts/documents in which the 2 words occur separately. Do they occur with the same set of context words? This can also be examined by looking at the vectors for the identity term and the highly associated other term in the term-context matrix. Do these share high values in certain dimensions that correspond to certain context words? Provide selected results and discuss findings in the report. Do you see evidence for representational harms (see below) learned by a bag-of-words model of this SNLI corpus? If so, which type do you see? Provide examples that support your conclusions. If you don’t find any potential**

harms, provide examples of what you examined and how you interpreted those associations.

  - See results below.

Something I want to comment on about all of the highlighted words–after doing nested searches on these words as the new target word, I feel that the world selection of the top 20 word ranking would have an overall positive or negative sentiment, but I didn't really know if this made sense or not.

#1 - Content Window Size = 4, word = "african"

```
The 10 most similar words to "african" using cosine-similarity on PPMI matrix are:
1: african; 1
2: american; 0.27961358020265825
3: americans; 0.22964856154052038
4: native; 0.20001777205096627
5: tribe; 0.17387364232976665
6: eastern; 0.15860295676212643
7: indian; 0.15723687997574787
8: traditional; 0.15212150893034848
9: tribal; 0.15186803675499616
10: rakish; 0.15116070365943723
11: asian; 0.1437487746976832
12: villagers; 0.13385326581774593
13: latino; 0.13174474576317396
14: youngsters; 0.1306630736253248
15: signer; 0.13006385803554343
16: incorrectly; 0.12975868906551102
17: ghetto; 0.12622309429876832
18: islamic; 0.12608255774303245
19: pastels; 0.12282568937059557
20: slum; 0.12260978420803292
```

Social Biases:

- There are plenty of African regions that do not have an association with tribal/villager-like living.
  - Not all Africans live in tribes
  - Not all Africans are tribal
  - Not all Africans are villagers

- There are many regions of Africa that are not associated with ghettos or slums, and many members of the subsets of African descent (e.g. African Americans) that are not associated with ghettos or slums
  - Not all Africans live in ghetto
  - Not all Africans live in slum

- Although there are muslims in regions of Africa, the Islamic faith originated in the Middle East (West Asia) and is the second largest religion. Furthermore, christianity is very popular in Africa.
  - Not all Africans are muslims (Islamic).

- Upon evaluation of the ranked words for ‘tribe’, the number 4 word is ‘african’.

The 10 most similar words to "tribe" using cosine-similarity on PPMI matrix are:
1: tribe; 1
2: aboriginal; 0.25884655075179575
3: dreamcatcher; 0.20554050878993846
4: african; 0.17387364232976665

The following 6 rows from the SNLI corpus is of 1st-order similarity to “african tribe”:

```
3564,1674493314.jpg#0,An older woman dressed in the garments of an African tribe has her right hand in a cracked brown pot and holds her left hand over another brown pot.
```

```
14270,3202804141.jpg#3,People from an African tribe are gathered outside.
```

```
18746,4418969015.jpg#0,Appears to be an African tribe or family posing for a picture in the grasslands with a dog in the corner.
```

```
41873,3736366270.jpg#2,A man of some native African tribe looking off into the distance.
```

```
87639,104824673.jpg#2,A member of an African tribe is watching the camera intently in tribal dress.
```

```
97811,4418969015.jpg#3,An African tribe is standing in their garden with the forest in the background.
```

```
143703,3202802351.jpg#3,Members of an African tribe are gathered in front of their huts.
```

```
574936,1147391743.jpg#4,This African tribe's women gather together to share stories about their husbands.
```

For comparison of frequency, there is one row for “american indian tribe”:

```
90115,5614228719.jpg#1,A member of an American Indian tribe plays some native music on a wind instrument at a local show.
```

And there are zero rows for “chinese tribe”, “american tribe”, “indian tribe” (exact), “asian tribe”, “australian tribe”, and “european tribe”.

The following 2 rows from the SNLI corpus is of 1st-order similarity to “african” and “tribal”:

```
101406,3203645080.jpg#3,African people in tribal wear in a desert.
```

```
126518,3203653804.jpg#3,A tribal African man with an assault rifle stands next to a
tribal African woman.
```

The following row from the SNLI corpus is of 1st-order similarity to “african” and “ghetto”:

```
437983,4944653840.jpg#0,An African woman is standing in a ghetto neighborhood waiting
for her husband.
```

After exclusively searching for the word “ghetto”, the only other specified race/ethnicity for an individual in the same sentence as ghetto (any other reference to an individual is not specified):

```
161772,4888308769.jpg#0,A black male walks through the ghetto.
```

#3 - Context Window Size = 4, word = ‘hate’

```
The 10 most similar words to "hate" using cosine-similarity on PPMI matrix are:
1: hate; 1
2: keg; 0.15956092158688961
3: panhandling; 0.1466378568942418
4: unbuttons; 0.14380086476594667
5: presume; 0.14072996041388142
6: untattooed; 0.13775975002103713
7: pleasure; 0.13691689225533354
8: shall; 0.13650603006220685
9: tenderloins; 0.13607712842936848
10: tearfully; 0.12779735228415956
11: withe; 0.12653877878194475
12: quiznos; 0.12606721318262948
13: hereford; 0.12318720957745322
14: chaps; 0.1229941535209067
15: scruffy; 0.1202474507389586
16: pod; 0.12014341085574176
17: assigned; 0.11980473606983844
18: maxi; 0.11955446659601798
19: nosy; 0.11825628627454154
20: dessed; 0.11743000277147853
```

Social Biases:

- The third ranked word for ‘hate’ is panhandling, which implies a hatred for panhandling and homelessness/joblessness.
    - We should never hate other humans due to their financial+residential status.

#4 - Context Window Size = 4, word = ‘adult’

```
The 10 most similar words to "adult" using cosine-similarity on PPMI matrix are:
1: adult; 1
2: males; 0.13413007712771963
3: young; 0.13029555653353608
4: insulated; 0.12633549281238576
5: father; 0.12424473432915928
6: swimsuits; 0.11823178578579563
7: toddler; 0.11258460982565821
8: small; 0.10899564711661469
9: imitates; 0.10836508257678601
10: adults; 0.10699526563938955
11: administering; 0.10683551283257031
12: boys; 0.10614150697015834
13: child; 0.10479373730241903
14: older; 0.10399574977701387
15: alternate; 0.10390446987183943
16: indoors; 0.10288002270476315
17: sad; 0.10069346360299081
18: children; 0.100519856382565
19: little; 0.099937843490851
20: girls; 0.0997166374579741
```

Social Biases:

- There are three references to the male gender, but there is only 1 reference to the female gender
    - Men and women can both be adults, but the data set is underrepresenting females in this case.
        - Being considered an adult holds value, and viewing a true adult as a child (as an example) is not good.

#5 - Context Window Size = 4, word = “president”

```
The 10 most similar words to "president" using cosine-similarity on PPMI matrix are:
1: president; 1
2: yes; 0.23326167715310364
3: clamour; 0.18540591728851807
4: gardner; 0.17045211470238675
5: unpopular; 0.16716946352187034
6: geography; 0.15462750530207603
7: mr; 0.15085143387313926
8: poland; 0.14778419469560844
9: custody; 0.144416099639449
10: crossdressing; 0.14383660849463564
11: let; 0.13663912761761998
12: vice; 0.13592123073069895
13: smugglers; 0.13461552046310876
14: puzzle; 0.1327726163482671
15: goodbyes; 0.1294787653621543
16: honest; 0.1271620392805879
17: babys; 0.12044079855942602
18: relationship; 0.11567475025677765
19: tartuffe; 0.11370245304857352
20: principal; 0.10732592875784752
```

Social Biases:

- There is only one reference to the male gender, but there are zero references to the female gender
    - Men and women may both become the president, but female leaders seem to be underrepresented in the data set.

#6 - Context Window Size = 4, word = “russian”

```
The 10 most similar words to "russian" using cosine-similarity on PPMI matrix are:
1: russian; 1
2: trusting; 0.18163153138349364
3: spy; 0.1760994951058914
4: benefactor; 0.14897166367392634
5: japan; 0.14757326337647247
6: spiking; 0.14604299482187022
7: usa; 0.14221916197369222
8: gps; 0.14100744854774638
9: draped; 0.1373630220241917
10: celegrate; 0.13655813630881108
11: comboy; 0.12845674047280065
12: confection; 0.12160857036765882
13: preserves; 0.11777920810482212
14: layup; 0.11740305059127387
15: forbidden; 0.11510016789355249
16: guiter; 0.1149874937328812
17: salutes; 0.11298001187010764
18: photographing; 0.1129740747553869
19: assists; 0.11190018402718294
20: preaching; 0.10903521060206234
```

Social Biases:

- The third ranked word for “russian” is “spy”, which implies a stereotype that Russians are associated with spying, but this is almost entirely false when considering the typical Russian (I would assume this is a residual stereotype in the USA post-cold war). Although Russia certainly has a highly-regarded spy force, I would argue that all first-world countries participate in global espionage.

- Upon evaluation of the ranked words for ‘spy’, the number 7 word is ‘russian’.

```
The 10 most similar words to "spy" using cosine-similarity on PPMI matrix are:
1: spy; 1
2: agent; 0.2330098439630539
3: sealing; 0.2236134274482956
4: recored; 0.20737769845374243
5: trusting; 0.19021760034622548
6: audiobook; 0.17790278280516536
7: russian; 0.1760994951058914
```

The following row from the SNLI corpus is of 1st-order similarity to “russian spy”:

```
241423,4851249911.jpg#2,The man is secretly a Russian spy watching the womens' every move.
```

As I mentioned above, this statement is implying a stereotype that Russians tend to be spies.

After exclusively searching for the word “russian”, the following row from the SNLI corpus was found that also supports this stereotype:

```
249311,6889203488.jpg#0,They are being watched by Russian spies.
```

This is unrelated to the point I have been trying to make, but here is another row that supports the stereotype that Russian individuals are always drinking vodka:

```
499550,7979042317.jpg#3,Three Russian men are drinking vodka.
```

After exclusively searching for the word “spy”, the following row from the SNLI corpus was found that also supports this stereotype:

```
183573,4389771657.jpg#1,a group of girls spy in soviet russia
```

By the way, there are zero row results in the SNLI corpus for “american spy”, “chinese spy”, “european spy”, “french spy”, and “indian spy”.