1. Assignment1 Report

Name: 潘郁婷

Institution (school): NTHU

Student ID: 113078506

Platform (Colab/Kaggle/Local): Local

Python version: 3.9.6

Operating system: MacOS (Apple Silicon M2)

CPU: Apple M2

GPU requirement: None

1. Which embedding model do you use? What are the pre-processing steps? What are the hyperparameter settings? (5%)

Answer:

* Embedding Model: Word2Vec
* Pre-processing Steps:
  + - * 1. Tokenization and Lowercasing: Each line is split into individual words (tokens), converted to lowercase, and punctuation is removed.
        2. Stopword and Non-English Word Removal: Common stop words (e.g., the, and, is) and non-alphabetic tokens are excluded to reduce noise.
        3. Lemmatization: Words are reduced to their base form to normalize variations.
        4. Yield Cleaned Tokens: Only non-empty token lists are kept and passed forward for training.
    - Hyperparameter Settings:
      1. vector\_size=100
      2. window=5
      3. min\_count=5
      4. workers=4
      5. sg=1
      6. epochs=5

1. What will the performance be like if you sample 5%, 10% and 20% of wiki text in TODO4? (10%, 3% for each)

Answer:

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **20%-Sample** | **10%-Sample** | **5%-Sample** |
| Vocab size | 1,054,999 | 678,669 | 442,848 |
| Semantic Accuracy | 60.7960% | 60.2886% | 58.5410% |
| Syntactic Accuracy | 33.8361% | 33.1803% | 31.6253% |

|  |  |  |  |
| --- | --- | --- | --- |
| **SubCategory** | **20%-Sample** | **10%-Sample** | **5%-Sample** |
| **capital-common-countries** | 92.2925% | 92.8854% | 90.5138% |
| **capital-world** | 78.6030% | 76.2157% | 74.5137% |
| **currency** | 17.6674% | 16.0508% | 11.5473% |
| **city-in-state** | 39.6838% | 42.6429% | 42.6429% |
| **family** | 46.8379% | 47.0356% | 41.6996% |
| **gram1-adjective-to-adverb** | 16.4315% | 16.4315% | 14.9194% |
| **gram2-opposite** | 14.1626% | 13.4236% | 11.3300% |
| **gram3-comparative** | 41.3664% | 41.3664% | 38.6637% |
| **gram4-superlative** | 22.9055% | 19.6970% | 18.0927% |
| **gram5-present-participle** | 36.4583% | 33.5227% | 31.6288% |
| **gram6-nationality-adjective** | 87.2420% | 87.4922% | 86.9919% |
| **gram7-past-tense** | 42.0513% | 41.7949% | 39.7436% |
| **gram8-plural** | 0.0% | 0.0% | 0.0% |
| **gram9-plural-verbs** | 10.3448% | 10.6896% | 8.3908% |

Using larger Wikipedia samples improves both vocabulary size and model performance. Semantic accuracy rises from 58.54% (5%) to 60.80% (20%), and syntactic accuracy from 31.63% to 33.84%, though gains diminish with more data. Strong categories like ‘capital-common-countries’ stay stable, while weak ones such as ‘currency’ and ‘plural’ remain challenging regardless of size.

1. What is the performance for different categories or sub-categories when trained on different corpora? (15%)

3.1 Present your results. (5%)

Answer:

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **20% Yelp Review** | **20 Newsgroups** | **5%-Sample Wiki** |
| Vocab size | 67,353 | 17,615 | 442,848 |
| Semantic Accuracy | 7.3965% | 0.8005% | 58.5410% |
| Syntactic Accuracy | 25.4333% | 2.6323% | 31.6253% |

|  |  |  |  |
| --- | --- | --- | --- |
| **SubCategory** | **20% Yelp Review** | **20 Newsgroups** | **5%-Sample Wiki** |
| **capital-common-countries** | 19.5652% | 3.1621% | 90.5138% |
| **capital-world** | 2.4315% | 0.3536% | 74.5137% |
| **currency** | 0.0% | 0.0% | 11.5473% |
| **city-in-state** | 11.5120% | 1.1350% | 42.6429% |
| **family** | 32.2134% | 2.1739% | 41.6996% |
| **gram1-adjective-to-adverb** | 14.8185% | 0.6048% | 14.9194% |
| **gram2-opposite** | 8.8670% | 0.0% | 11.3300% |
| **gram3-comparative** | 59.5345% | 7.6577% | 38.6637% |
| **gram4-superlative** | 27.2727% | 0.7130% | 18.0927% |
| **gram5-present-participle** | 38.9205% | 6.723% | 31.6288% |
| **gram6-nationality-adjective** | 23.8899% | 1.06316 % | 86.9919% |
| **gram7-past-tense** | 8.2051% | 4.9359% | 39.7436% |
| **gram8-plural** | 0.0% | 0.0% | 0.0% |
| **gram9-plural-verbs** | 0.9195% | 0.0% | 8.3908% |

3.2 Introduce the corpus you selected and explain the differences between the Wikipedia corpus and your corpus. (including data size, topic difference, structural difference … ) (5%)

Answer:

* 1. 5%-Sample Wikipedia
  2. [The 20 Newsgroups data set](http://qwone.com/~jason/20Newsgroups/)
  3. [20% Sample of Yelp Review](https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_review.json)

Introduction:

|  |  |  |  |
| --- | --- | --- | --- |
| **Corpus** | **vocabular size** | **topic difference** | **Structural difference** |
| 5%-Sample Wikipedia | 442,848 | Wide-ranging factual topics across domains | Plain text articles with hyperlinks, section headings, and markup structure |
| 20 Newsgroups | 17,615 | 20 discussion categories (politics, sports, tech, religion, etc.) | Plain text documents, each file = one post, labeled by newsgroup category |
| Yelp Reviews (20% Sample) | 67,353 | Consumer reviews about restaurants and businesses | JSON format; each record contains fields like review\_id, user\_id, stars, text |

3.3 Explain why the accuracy increases or decreases. (5%)

Answer:

Accuracy decreases when training on smaller or noisier corpora because the model has fewer examples to learn word relationships, leading to limited vocabulary coverage. Domain-specific texts like Yelp, Newsgroups lack the diversity and consistency of Wikipedia, so the learned embeddings capture narrower contexts and perform poorly on broad analogy tasks.

The results in 3.1 show that the Wikipedia sample achieves the highest semantic and syntactic accuracy, benefiting from its large size and well-structured factual content. In contrast, Yelp reviews and 20 Newsgroups perform much worse, especially in semantic tasks, probably due to their smaller size and noisier, domain-specific language.

Intriguingly, the Yelp corpus outperforms Wikipedia and Newsgroups on categories like ‘opposites’, ‘comparatives’, ‘superlatives’, and ‘participles’. This is likely because reviews often use such grammatical forms in everyday expressions, giving the model richer patterns to learn despite its weaker overall semantic accuracy.

In addition, the Yelp corpus performs poorly on the ‘currency’ category, this is probably because reviews rarely mention currency names or exchange rates.

1. Select a few words and use their embeddings to retrieve the five most similar words and present the results. What do you observe? (10%)

Answer:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **20%-Sample TOP-5** | **10%-Sample TOP-5** | **5%-Sample TOP-5** |
| king | prince (similarity: 0.7944)  queen (similarity: 0.7449)  monarch(similarity: 0.7275)  throne (similarity: 0.7272)  reign (similarity: 0.7189) | prince (similarity: 0.7524)  queen (similarity: 0.7109)  uncrowned (similarity:0.7100)  nangklao (similarity: 0.7069)  throne (similarity: 0.7022) | monarch(similarity:0.7667)  queen (similarity: 0.7322)  throne (similarity: 0.7208)  vajiravudh(similarity:0.7176)  abdicates(similarity: 0.7157) |
| queen | princess (similarity: 0.7613)  king (similarity: 0.7449)  soheon (similarity: 0.7221)  iiprince (similarity: 0.7217)  inwon (similarity: 0.7183) | inmok (similarity: 0.7259)  princess (similarity: 0.7161)  king (similarity: 0.7109)  kamāmalu (similarity: 0.7023)  iiprince (similarity: 0.6975) | king (similarity: 0.7322)  kamāmalu(similarity:0.7270)  princess (similarity: 0.7227)  inmok (similarity: 0.7070)  sirikit (similarity: 0.7005) |
| man | person (similarity: 0.7184)  young (similarity: 0.7162)  boy (similarity: 0.6772)  girl (similarity: 0.6659)  way (similarity: 0.6571) | young (similarity: 0.6978)  person (similarity: 0.6807)  boy (similarity: 0.6663)  특집 similarity: 0.6628)  uniland (similarity: 0.6489) | young (similarity: 0.7079)  minsu (similarity: 0.6802)  naksu (similarity: 0.6777)  quanxi (similarity: 0.6764)  bakwas (similarity: 0.6682) |
| woman | men (similarity: 0.8098)  individual(similarity:0.7146)  girl (similarity: 0.7047)  ghart (similarity: 0.7036)  bialova (similarity: 0.6752) | men (similarity: 0.8121)  iwsf (similarity: 0.6841)  girl (similarity: 0.6746)  medalled (similarity: 0.6703)  histofina (similarity: 0.6692) | men (similarity: 0.8138)  usrowing (similarity: 0.6958)  triathletes(similarity: 0.6700)  escanellas(similarity:0.6640)  individual(similarity:0.6602) |
| Paris | marseille(similarity: 0.8156)  bagnolet (similarity: 0.7928)  france (similarity: 0.7887)  brussels (similarity: 0.7788)  maratier (similarity: 0.7738) | marseille (similarity: 0.8008)  vivrel (similarity: 0.7931)  france (similarity: 0.7841)  beaubourg(similarity: 0.7841)  villeurbanne(similarity:0.7822) | marseille (similarity: 0.8033)  billancourt(similarity:0.7990)  france (similarity: 0.7968)  créteil (similarity: 0.7818)  bercy(similarity: 0.7802) |
| China | taiwan (similarity: 0.8725)  guangdong(similarity:0.8262)  beijing (similarity: 0.8040)  guangxi (similarity: 0.7972)  shanghai (similarity: 0.7953) | taiwan (similarity: 0.8745)  guangdong (similarity:0.8281)  tianjin (similarity: 0.8060)  guangxi (similarity: 0.8047)  chinese (similarity: 0.7992) | taiwan (similarity: 0.8753)  guangdong(similarity:0.8151)  shenzhen (similarity: 0.8080)  sichuan (similarity: 0.8037)  guangxi (similarity: 0.7993) |

|  |  |  |
| --- | --- | --- |
|  | **YELP Review 20% Sample TOP-5** | **20 Newsgroups TOP-5** |
| king | prussia (similarity: 0.6454)  meches (similarity: 0.5401)  boudoin (similarity: 0.5346)  osiris (similarity: 0.5247)  caluda (similarity: 0.5143) | murphy (similarity: 0.8005)  rodney (similarity: 0.7827)  brendan (similarity: 0.7785)  joseph (similarity: 0.7759)  blackhawks(similarity:0.7754) |
| queen | sheba (similarity: 0.5712)  pullout (similarity: 0.5263)  sheeba (similarity: 0.5248)  king (similarity: 0.4998)  reseda (similarity: 0.4979) | blew (similarity: 0.8684)  bronx (similarity: 0.8578)  snapped (similarity: 0.7952)  panicking (similarity: 0.7908)  garrett (similarity: 0.7895) |
| man | gentleman(similarity:0.7277)  lady (similarity: 0.6875)  guy (similarity: 0.6865)  woman (similarity: 0.6841)  teki (similarity: 0.6158) | straw (similarity: 0.6790)  thou (similarity: 0.6668)  tale (similarity: 0.6627)  behold (similarity: 0.6550)  hath (similarity: 0.6532) |
| woman | lady (similarity: 0.8315)  girl (similarity: 0.6979)  gentleman(similarity:0.6859)  man (similarity: 0.6841)  female (similarity: 0.6654) | pregnant (similarity: 0.8082)  elderly (similarity: 0.8066)  corpse (similarity: 0.7865)  girl (similarity: 0.7810)  dressed (similarity: 0.7770) |
| Paris | france (similarity: 0.6921)  arrondissement(similarity:0.6835)  parisian (similarity: 0.6242)  relais (similarity: 0.6171)  ladurée (similarity: 0.6159) | aiu (similarity: 0.9167)  emeritus (similarity: 0.9003)  lausanne (similarity: 0.8976)  shaw (similarity: 0.8932)  hairenik (similarity: 0.8927) |
| China | chinese (similarity: 0.7033)  yuan(similarity: 0.6219)  chine (similarity: 0.5889)  toishan (similarity: 0.5875)  guangzhou (similarity: 0.5867) | fatah (similarity: 0.9053)  picket (similarity: 0.9048)  falkland (similarity: 0.9018)  libya (similarity: 0.9005)  sultan (similarity: 0.8988) |

1. Wikipedia (5% / 10% / 20% samples) produces the most meaningful semantic neighbors. For example, ‘king’ is strongly linked to ‘queen’, ‘prince’, ‘throne’, and ‘Paris’ is linked to other French cities like ‘Marseille’ and ‘Brussels’. This reflects Wikipedia’s large, factual, and diverse coverage.
2. Yelp reviews capture everyday associations but lack semantic depth. For instance, ‘man’ is close to ‘gentleman’,’ lady’, ‘guy’, ‘woman’, showing its conversational nature, but ‘Paris’ is associated with terms like ‘ladurée’(a pastry shop), which reflects Yelp’s food and business bias.
3. 20 Newsgroups yields noisier and less semantically consistent neighbors. Words like ‘queen’ map to verbs ‘blew’, ‘snapped’ and ‘China’ to unrelated political terms ‘fatah’, ‘falkland’, because the corpus is much smaller, domain-specific, and discussive.
4. Anything that can strengthen your report. (5%)

Answer:

To strengthen the report, I wanted to do an error analysis combined with a cross-corpus comparison because accuracy numbers alone don’t really show the full picture. By looking at errors and comparing across corpora, it’s easier to see what each dataset is good at, what it struggles with, and which applications it might fit best.

|  |  |  |  |
| --- | --- | --- | --- |
| **Corpus** | **Strength** | **Weakness** | **Best suited for** |
| Wikipedia (Sampled) | Large and diverse; learns solid semantic relationships | Weak on rare syntactic patterns (e.g., plurals) | General NLP tasks, knowledge extraction, baseline models |
| Yelp Reviews (20% Sample) | Everyday language; good at comparatives, superlatives, verb forms | Little coverage of currency or world facts | Sentiment analysis, opinion mining, customer review studies |
| 20 Newsgroups | Topic-labeled discussions; captures informal writing styles | Small and noisy; odd links (e.g., ‘queen’ → ‘blew’) | Topic classification, conversational or stylistic analysis |