**Assignment1 Report**

Name:鄭慈昱

Institution (school):國立清華大學

Student ID:113078505

Platform (Colab/Kaggle/Local): Local

Python version: 3.9.6

Operating system: macOS(Apple Silicon M4)

CPU: Apple M4

GPU requirement: None

#AI Usage

As a beginner in Natural Language Processing (NLP), I relied on **ChatGPT** to gradually clarify my questions and guide me step by step. Through prompt engineering, I was able to:

* Break down complex tasks into manageable steps.
* Refine my report content with clearer structure and academic wording.
* Receive explanations and examples that helped me understand both technical implementation and result interpretation.

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

Answer: I used **Word2Vec** from Gensim.

* **Pre-processing:**
  + Tokenization using simple\_preprocess
  + Sentence segmentation by punctuation.
  + Removal of empty lines
  + Normalization of words: lowercasing, accent removal, symbol cleaning, and checking alternative forms (e.g., - / \_, removing plural *s*, removing apostrophes) to increase vocabulary match
* **Hyperparameters:**
  + Vector size: 200
  + Window size: 5
  + min\_count: 5
  + sg =1 for skip-gram
  + 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:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 5% Sample | 10% Sample | 20% Sample |
| Vocabulary size | 442,848 | 688,697 | 1,054,999 |
| Overall Accuracy | 0.4456 | 0.4403 | 0.6389 |
| Semantic Accuracy | 0.5700 | 0.5532 | 0.7430 |
| Syntactic Accuracy | 0.3416 | 0.3459 | 0.5514 |

|  |  |  |  |
| --- | --- | --- | --- |
| Sub-category | 5% Sample | 10% Sample | 20% Sample |
| **capital-common-countries** | 0.8676 | 0.7925 | 0.9506 |
| **capital-world** | 0.6452 | 0.6494 | 0.8888 |
| **currency** | 0.0370 | 0.0462 | 0.1524 |
| **city-in-state** | 0.5428 | 0.4966 | 0.6178 |
| **family** | 0.6443 | 0.5968 | 0.8538 |
| **gram1-adjective-to-adverb** | 0.1159 | 0.1341 | 0.2198 |
| **gram2-opposite** | 0.0973 | 0.1342 | 0.2302 |
| **gram3-comparative** | 0.3378 | 0.3431 | 0.6749 |
| **gram4-superlative** | 0.1212 | 0.0720 | 0.4081 |
| **gram5-present-participle** | 0.2936 | 0.3125 | 0.4820 |
| **gram6-nationality-adjective** | 0.7423 | 0.7674 | 0.8837 |
| **gram7-past-tense** | 0.3731 | 0.3455 | 0.4859 |
| **gram8-plural** | 0.3619 | 0.3455 | 0.6667 |
| **gram9-plural-verbs** | 0.3345 | 0.3897 | 0.6080 |

**Conclusion:**

Although the 10% subset is larger than 5%, its overall and semantic accuracy are slightly lower, likely due to sampling variance and added noise. However, the syntactic accuracy of 10% is marginally higher, suggesting that a larger corpus provides more structural linguistic patterns, even if semantic relations are less emphasized. Overall, the 20% corpus achieves the best balance.

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

3.1 Present your results. (5%)

Answer:

**This section is compared with four categories:**

* **20% sample of Wikipedia articles**
* **HuggingFace - TimSchopf medical\_abstracts dataset. [**[link](https://huggingface.co/datasets/TimSchopf/medical_abstracts?utm_source=chatgpt.com)**]**
* **HuggingFace – NewsGroup [**[link](https://huggingface.co/datasets/fancyzhx/ag_news)**]**
* **HuggingFace – DBPedia [**[link](https://huggingface.co/datasets/fancyzhx/dbpedia_14)**]**

**The following table summarizes the analogy task results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Wikipedia (20%) | Medical abstracts | NewsGroup | DBpedia |
| Overall Accuracy | 0.6389 | 0.0643 | 0.1663 | 0.2799 |
| Semantic Accuracy | 0.7430 | 0.2645 | 0.2072 | 0.3827 |
| Syntactic Accuracy | 0.5514 | 0.0496 | 0.1443 | 0.1899 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sub-category | Wikipedia (20%) | Medical abstracts | NewsGroup | DBpedia |
| capital-common-countries | 0.9506 | 0.0000 | 0.5584 | 0.5830 |
| capital-world | 0.8888 | 0.3333 | 0.3303 | 0.3704 |
| currency | 0.1524 | 0.0000 | 0.0588 | 0.0000 |
| city-in-state | 0.6178 | 0.0278 | 0.0414 | 0.3895 |
| family | 0.8538 | 0.3545 | 0.1083 | 0.4675 |
| gram1-adjective-to-adverb | 0.2439 | 0.0105 | 0.0053 | 0.0410 |
| gram2-opposite | 0.2007 | 0.0179 | 0.0065 | 0.0556 |
| gram3-comparative | 0.7913 | 0.0850 | 0.1254 | 0.1044 |
| gram4-superlative | 0.5428 | 0.0985 | 0.0897 | 0.0851 |
| gram5-present-participle | 0.6951 | 0.0125 | 0.0713 | 0.0527 |
| gram6-nationality-adjective | 0.8787 | 0.0097 | 0.4873 | 0.7086 |
| gram7-past-tense | 0.5545 | 0.0234 | 0.0972 | 0.0865 |
| gram8-plural | 0.7200 | 0.0440 | 0.0632 | 0.1143 |
| gram9-plural-verbs | 0.5839 | 0.3194 | 0.0983 | 0.0688 |

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:

|  |  |  |  |
| --- | --- | --- | --- |
| Corpus | Data Size | Topic Domain | Structure Difference |
| Wikipedia (20%) | 1,054,999 | General encyclopedic | Formal, narrative,  balanced vocabulary |
| Medical Abstracts | 11,550 | Biomedical research | Highly technical,  dense terminology |
| Newsgroup | 120,000 | Mixed (politics, sports, etc.) | Semi-formal,  sometimes colloquial |
| DBpedia | 560,000 | Structured encyclopedic data | Formal, concise,fragmentary |

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

Answer:

* **Wikipedia (20%) achieved the best accuracy (Overall Accuracy = 0.6389). Its scale and balanced coverage supported strong semantic tasks (*capital-common-countries, nationality-adjective*) and relatively stable syntactic performance, though complex transformations (*adjective-to-adverb*) stayed weaker.**
* **Medical abstracts performed poorly (Overall Accuracy = 0.0643). The narrow biomedical focus left most semantic categories absent (*currency, capital*), with only moderate results in *family* terms due to genetic/clinical usage. Its technical, noun-heavy style also restricted syntactic diversity.**
* **Newsgroup showed modest gains (Overall Accuracy = 0.1663). Informal discussions captured everyday syntax (*plural, past tense*), but encyclopedic tasks (*city-in-state, capital-world*) remained low. Very weak results in *adjective-to-adverb* and *opposites* reflected its colloquial style.**
* **DBpedia outperformed Medical and Newsgroup (Overall Accuracy = 0.2799). Its structured format provided strong entity coverage (*capital-common-countries, nationality-adjective*), but the fragmentary infobox style lacked grammatical variation, lowering *gram1–5* scores.**

**Conclusion:**  
Accuracy trends follow both corpus size and thematic scope. General encyclopedic text (Wikipedia) yields the most balanced embeddings, domain-limited or informal text (Medical, Newsgroup) miss many categories, and structured data (DBpedia) strengthens entity accuracy while weakening grammatical relations.

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:

|  |  |  |  |
| --- | --- | --- | --- |
| Word | 5% Sample Top-5 | 10% Sample Top-5 | 20% Sample Top-5 |
| king | queen, kingdom, royal, prince, henry | george, queen, royal, henry, event | queen, prince, nangklao, phuthalotla, monarch |
| queen | king, princess, lagebaton, auscheer, sebele | king, princess, chadabin, elizabeth, mehones | king, princess, jangnyeol, cheorin, inmok |
| man | old, women, men, girl, way | show, old, live, very, instead | woman, girl, boy, person, old |
| woman | wonder, metamorpho, fcbd, girl, cbldf | wonder, girl, analist, endsong, funnest | man, girl, person, 俗女養成記, ohitika |
| city | but, only, two, both, them | but, only, two, both, being | town, area, cities, citythe, district |
| country | music, town, john, august, final | them, including, second, third, however | nation, startorder, countrytime, korrika, warrabri |

**Observations:**

* **Smaller corpora (5% and 10%)** produced many noisy or irrelevant neighbors (e.g., lagebaton, cbldf, funnest). Function words such as but, only, and them also appeared, showing insufficient semantic learning.
* **10% vs 5%:** While the 10% subset had more syntactic consistency (king → george, queen, royal), it still contained spurious tokens.
* **20% subset:** Results were more coherent and semantically meaningful. Gender terms clustered together (man ↔ woman, girl, boy), and geographic entities aligned better (city → town, district; country → nation). However, some noise persisted (e.g., transliterated names like nangklao, jangnyeol).

**Conclusion:**

Increasing corpus size reduces noise and enhances semantic clustering, but the presence of rare or foreign tokens indicates that even 20% sampling may not fully stabilize the embedding space.

1. Anything that can strengthen your report. (5%)

Answer:

To strengthen the analysis, I visualized word embeddings from different corpora using t-SNE. The results reveal distinct clustering patterns:

一張含有 文字, 螢幕擷取畫面, 圖表, 行 的圖片

AI 產生的內容可能不正確。

* **News** captures basic clusters for people and places, while topic-specific terms remain more dispersed.
* **Medical abstracts** group biomedical terms together but scatter general words, highlighting domain specialization.
* **DBpedia** clusters entity names (e.g., countries, companies) tightly, reflecting its structured, entity-focused content.
* **Wikipedia (20%)** shows the clearest separation, with people, geographic, and thematic words forming well-defined groups.

These visualizations confirm that corpus theme and style not only affect accuracy in analogy tasks but also shape the semantic space of embeddings, providing an intuitive complement to quantitative results.