A brief description of the corpus/corpora you selected for training and what preprocessing steps you took. Mention the total number of sentences.

I have selected **NLTK Brown corpus** for completing this assignment. The NLTK Brown corpus is a widely used corpus in natural language processing and linguistics. It is a collection of text from a variety of sources, including news articles, fiction, and non-fiction texts, and it is categorized into different genres (e.g., news, fiction, religion).

Preprocessing steps I took was Sentence Segmentation and Lowercasing.

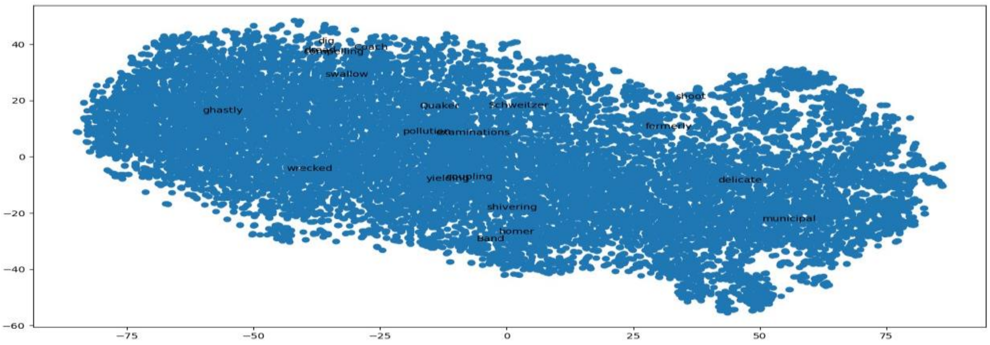
Total Number of Sentences: 57340

Mention in which two different ways you decided to get the word embeddings and why.

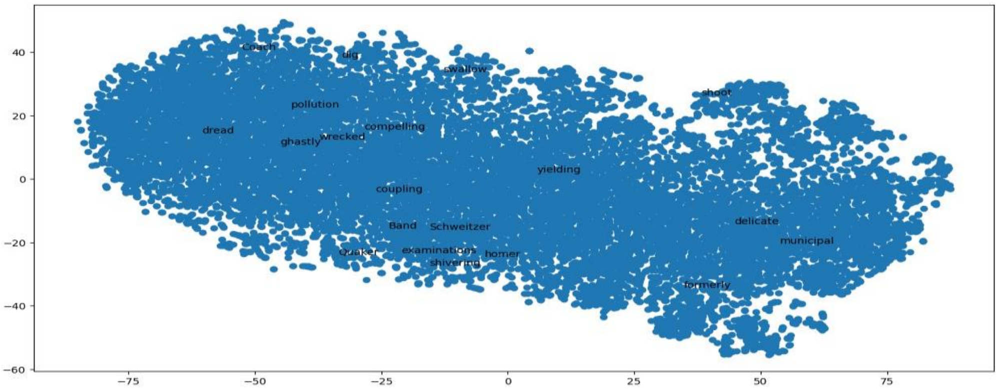
1. Continuous Bag of Words (CBOW): It can perform well when the training data has a large amount of context information and the vocabulary size is large.

2. Skip-Gram: Skip-Gram can capture more fine-grained semantic relationships between words. It is better at capturing nuances and can be useful when you want to obtain word embeddings that reflect the rich semantic information present in the data.

The two graphs of visualization from Task 2. Write a few comments about them.



CBOW: In the CBOW embedding visualization, words with similar semantic meanings tend to cluster together. This suggests that the CBOW model has successfully captured semantic relationships between words in the training data.



Skipgram: The Skip-Gram embedding visualization often shows a more scattered arrangement of words. This suggests that the Skip-Gram model focuses on capturing fine-grained relationships between words, allowing for greater flexibility in word representations.

The results of Task 3 for the three-word embedding in one table (correlation scores). Write comments on the results.

	PearsonRResult	SignificanceResult	
Google Embedding	(statistic=0.42055563497789983, pvalue=0.2596917638131105)	(statistic=0.48535989707831934, pvalue=0.18535381786420974)	0.0
CBOW	(statistic=0.4161130491613724, pvalue=0.26528670138358346)	SignificanceResult(statistic=0.5272012675161055, pvalue=0.14469933213386177)	0.0
SkipGram	(statistic=0.4223108969885318, pvalue=0.25749859646583395)	SignificanceResult(statistic=0.6276205565667923, pvalue=0.07036502278970762)	0.0

Google Embedding: Moderate correlation (0.4206), indicating a reasonable similarity to the human-judged scores.

CBOW: Similar to Google Embedding (0.4161), with a moderate correlation.

SkipGram: Higher correlation (0.4223), suggesting a slightly better alignment with human-judged similarity scores.

The results of Task 4 with some comments.

The 5 words are: municipal, shoot, delicate, dig, swallow

- **Google-Embedding:** [('municipality', 0.7065823674201965), ('municipalities', 0.6907703876495361), ('muncipal', 0.6503286957740784), ('Municipal', 0.6325821280479431), ('city', 0.622254490852356), ('manager_Gaster_Sharpley', 0.5772400498390198), ('munici_pal', 0.5740675926208496), ('Municipality', 0.5608386993408203), ('municpal', 0.5585163235664368), ('township', 0.5535756349563599)]

[('shooting', 0.6357938647270203), ('shoots', 0.621467113494873), ('Shoot', 0.6096479296684265), ('shot', 0.6021238565444946), ('shooters', 0.5309295058250427), ('Shooting', 0.5165857672691345), ('shots', 0.5091301202774048), ('shoot_skeet', 0.4836919605731964), ('kill', 0.48082584142684937), ('Aliodor_recalled', 0.44888991117477417)]

[('delicately', 0.6477687954902649), ('Delicate', 0.5852839946746826), ('fragile', 0.5694279670715332), ('utmost_delicacy', 0.5459215044975281), ('intricate', 0.5406070947647095), ('tricky', 0.5351399183273315), ('sensitive', 0.5144727826118469), ('delicately_balanced', 0.5091362595558167), ('thorny', 0.5041804313659668), ('complicated', 0.5028288960456848)]

[('digging', 0.7896488904953003), ('dug', 0.7848839163780212), ('excavate', 0.634793221950531), ('Dig', 0.6323593854904175), ('Digging', 0.6184645295143127), ('dig_deep', 0.6082916259765625), ('dig_deeper', 0.607262909412384), ('digging_deeper', 0.5682678818702698), ('bury', 0.548970639705658), ('unearth', 0.5485262870788574)]

[('swallowed', 0.6274389624595642), ('swallowing', 0.6222291588783264), ('bitterest_pill', 0.5339366793632507), ('chew', 0.520850419998169), ('Bitter_pill', 0.5149303674697876), ('swallows', 0.5145219564437866), ('gulp', 0.4934978187084198), ('Oral_sprays', 0.4860546886920929), ('digest', 0.4629243314266205), ('suck', 0.45565924048423767)]

- **CBOW:** [('processes', 0.9723106026649475), ('matching', 0.9716562628746033), ('existing', 0.9709226489067078), ('payments', 0.9702702760696411), ('biological', 0.968867838382721), ('solely', 0.9686510562896729), ('earnings', 0.9683205485343933), ('securing', 0.9681024551391602), ('storage', 0.967593789100647), ('applications', 0.9671738147735596)]

  [('settle', 0.9536823034286499), ('blame', 0.949759840965271), ('swing', 0.9415808320045471), ('catch', 0.9409534335136414), ('sit', 0.9370325803756714), ('rent', 0.9336565136909485), ('grow', 0.9330014586448669), ('jump', 0.9320572018623352), ('count', 0.9315732717514038), ('workshop', 0.9308232069015503)]

  [('vigorous', 0.9736443758010864), ('urbanization', 0.9684687852859497), ('bold', 0.965545654296875), ('Asian', 0.9636419415473938), ('copper', 0.9629600644111633), ('well-known', 0.9627700448036194), ('Styka', 0.9627517461776733), ('spectacular', 0.962187647819519), ('rhythms', 0.9619085192680359), ('fruit', 0.9609394073486328)]

  [('Him', 0.930173933506012), ('freely', 0.9227550625801086), ('attempted', 0.9200465083122253), ('cereal', 0.9172289371490479), ('blame', 0.9142351746559143), ('negotiate', 0.9131203889846802), ('strongly', 0.9130229949951172), ('please', 0.9121226072311401), ("'im", 0.9112053513526917), ('answers', 0.9110734462738037)]

  [('earnestly', 0.9468746781349182), ('cease', 0.9390227794647217), ('welcome', 0.9330333471298218), ('spark', 0.9325726628303528), ('assure', 0.930778980255127), ('save', 0.928370475769043), ('hoping', 0.9271546602249146), ("o'", 0.9260084629058838), ('express', 0.9247851967811584), ('remind', 0.9241346120834351)]

- **SkipGram:** [('governmental', 0.9365421533584595), ('purchasing', 0.9250916838645935), ('civic', 0.9249060750007629), ('unions', 0.9246639013290405), ('improvements', 0.9241287708282471), ('maintaining', 0.922650158405304), ('markets', 0.9198223948478699), ('cooperative', 0.9162508249282837), ('strengthening', 0.9161201119422913), ('commerce', 0.9150803089141846)]

  [('swing', 0.9414761662483215), ('snap', 0.9323044419288635), ('scream', 0.930915892124176), ('hang', 0.9266887307167053), ('jump', 0.9265965223312378), ('hide', 0.9237306118011475), ('Grandma', 0.9216713905334473), ('climb', 0.9174028038978577), ('ride', 0.9156705737113953), ('weep', 0.914165735244751)]

  [('rhythmic', 0.9458966851234436), ('shrewd', 0.9400585889816284), ('asset', 0.9331402778625488), ('delicacy', 0.9322227835655212), ('recurring', 0.9310033917427063), ('disintegration', 0.9308595061302185), ('functioning', 0.9305880665779114), ('age-old', 0.9301460385322571), ('plight', 0.929745614528656), ('stance', 0.9296683073043823)]

```
[('reckon', 0.9750247597694397), ('despise', 0.970294713973999),
('dodge', 0.9700610637664795), ('queens', 0.9686623215675354),
('grok', 0.9681969881057739), ('Personally', 0.9660098552703857),
('Were', 0.9653280377388), ('typing', 0.9640051126480103), ('breathe',
0.9633293747901917), ('cereal', 0.9614673852920532)]
```

```
[('stir', 0.967476487159729), ('execute', 0.9667884111404419), ('anew',
0.966356635093689), ('wink', 0.9645264744758606), ('wrongs',
0.9641938805580139), ('submitting', 0.964169979095459), ('barbell',
0.9639124870300293), ('Andrena', 0.9634842872619629), ('leisure',
0.9632096886634827), ('Matson', 0.9631139636039734)]
```

- Google Embeddings generally provide relevant and related words for the given words.
- CBOW results are less relevant for some words, indicating limitations in capturing specific word relationships.
- SkipGram performs well in capturing relevant words for the given words.