Solution

Part 1: Corpus Analysis

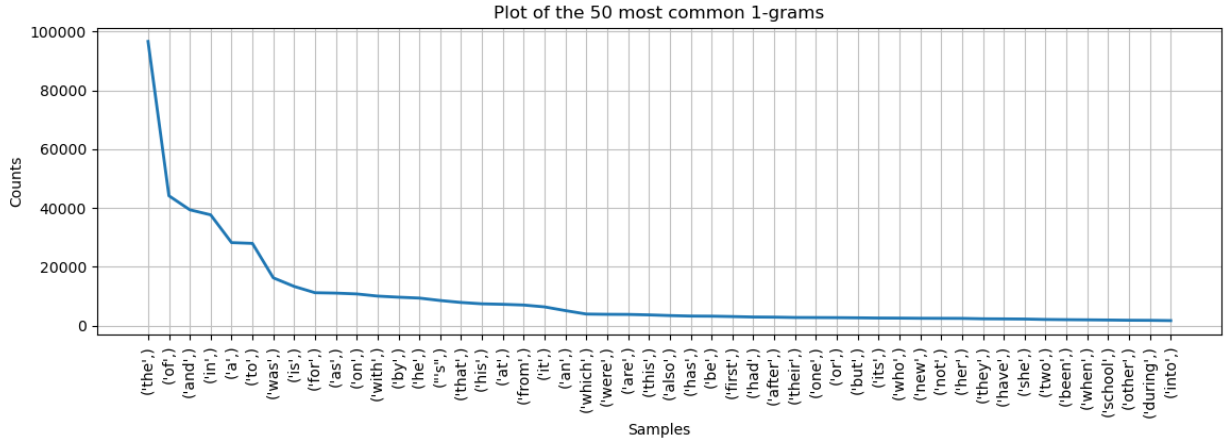
Q1. Unigram analysis:

(a) Mention the total unique unigrams present in the corpus

Total unique 1-grams : 75645

(b) Plot the distribution of the unigram frequencies

Top 50 most common words



(c) How many (most frequent) uni-grams are required to cover the 90% of the complete corpus.

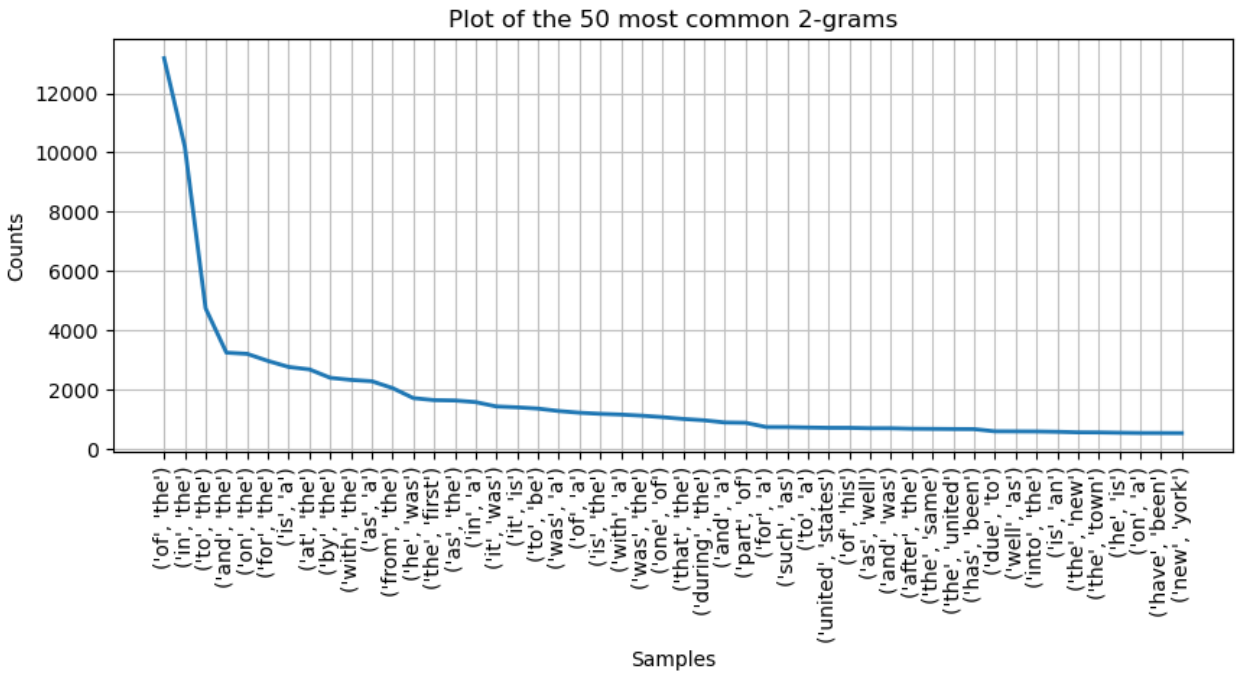
1-grams required to cover 90% of the complete corpus : 11714

Q2. Bigram analysis:

(a) Mention the total unique bigrams present in the corpus.

Total unique 2-grams : 581899

(b) Plot the distribution of the bigram frequencies.



(c) How many (most frequent) bi-grams are required to cover the 90% of the complete corpus.

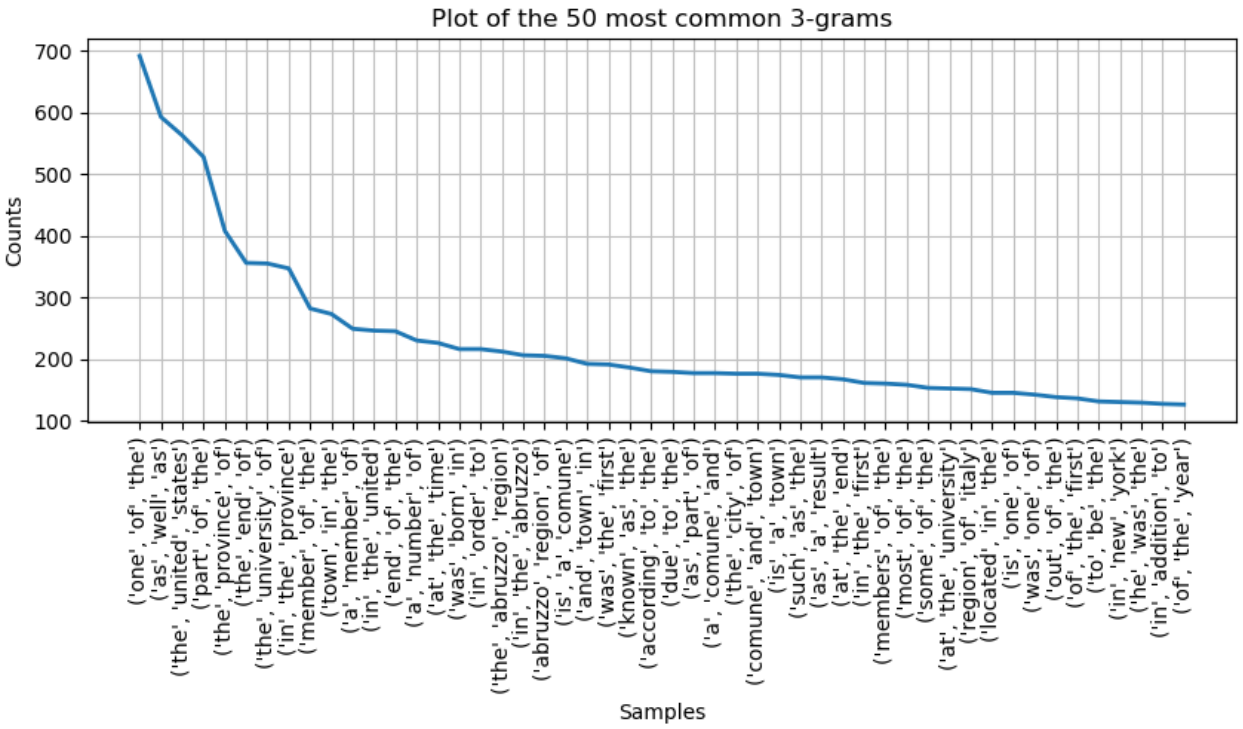
2-grams required to cover 90% of the complete corpus : 454207

Q3. Trigram analysis:

(a) Mention the total unique trigrams present in the corpus.

Total unique 3-grams : 1044041

(b) Plot the distribution of the trigram frequencies.



(c) How many (most frequent) tri-grams are required to cover the 90% of the complete corpus.

3-grams required to cover 90% of the complete corpus : 919608

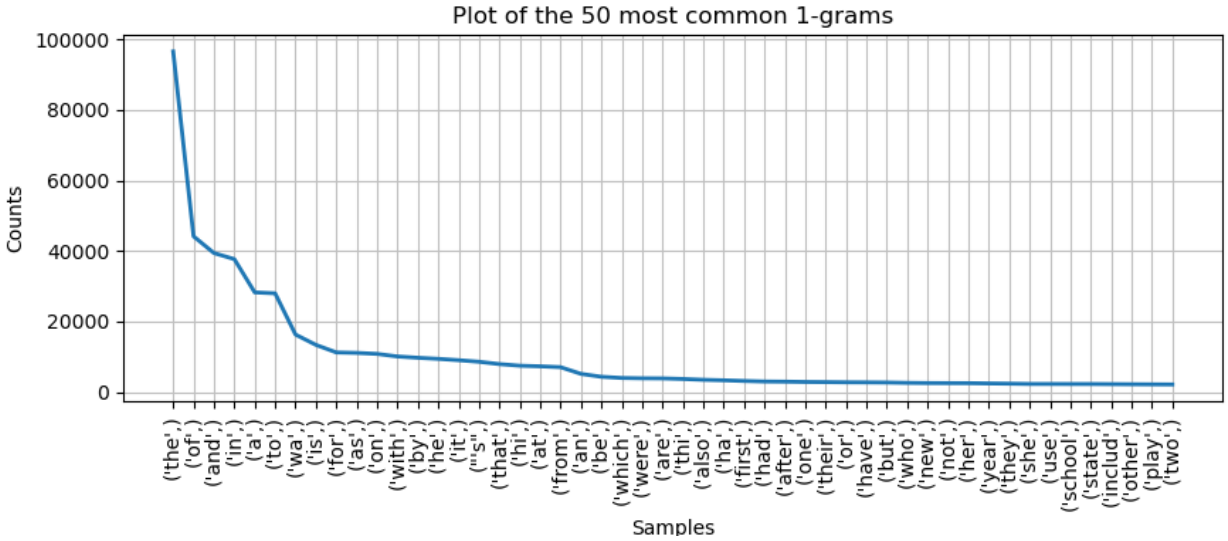
Q4. Repeat Q1, Q2, and Q3 after performing the stemming process on the tokens.

Unigram analysis:

(a) Mention the total unique unigrams present in the corpus.

Total unique 1-grams : 60644

(b) Plot the distribution of the unigram frequencies.



(c) How many (most frequent) uni-grams are required to cover the 90% of the complete corpus.

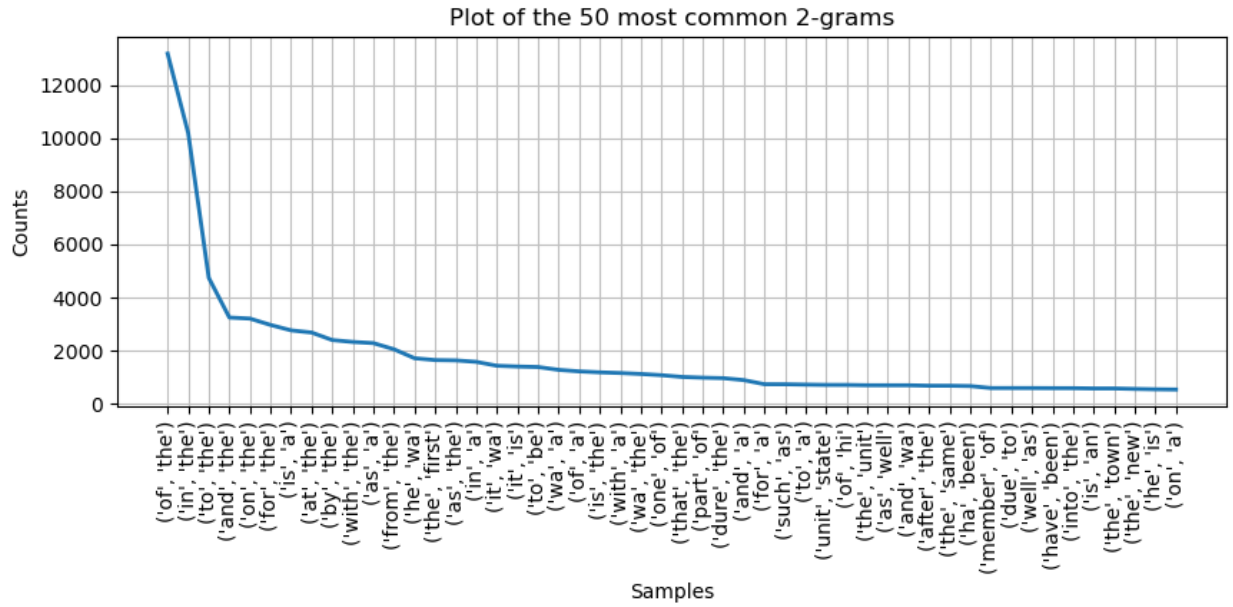
1-grams required to cover 90% of the complete corpus : 6605

Bigram analysis

(a) Mention the total unique bigrams present in the corpus.

Total unique 2-grams : 534154

(b) Plot the distribution of the bigram frequencies.



(c) How many (most frequent) bi-grams are required to cover the 90% of the complete corpus.

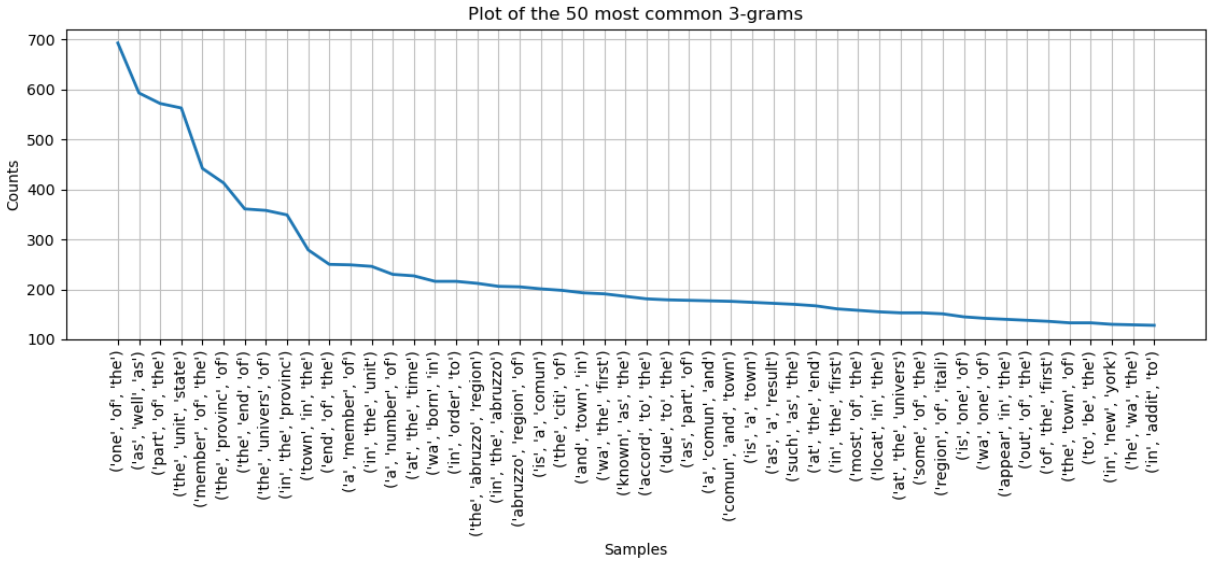
2-grams required to cover 90% of the complete corpus : 406462

Trigram analysis:

(a) Mention the total unique trigrams present in the corpus.

Total unique 3-grams : 1025690

(b) Plot the distribution of the trigram frequencies.



(c) How many (most frequent) tri-grams are required to cover the 90% of the complete corpus.

3-grams required to cover 90% of the complete corpus : 901257

Q5 Briefly summarize and discuss the frequency distributions obtained in Q1 to Q4. Do these distributions approximately follow Zipf's law?

Yes. It does follow zipfs law approximately especially for unigrams. We can see that in charts plotted for every n-gram case atleast for top 5 ngrams before it becomes constant

Q6. What library you used for tokenization and stemming? What were the underlying algorithms used by the library for these tasks?

I used nltk tokenization and nltk stemming libraries. Nltk stemmer uses Porter Stemmer Algorithm. NLTK Tokenizer uses TreebankWordTokenizer along with PunktSentenceTokenizer for the specified language

Q7. Report three examples based on your observation, where the tool used for tokenization did not tokenize the character sequence properly.

i) It ended up taking any word with double quotes like “ABC” as 3 different tokens. This can be solved by removing punctuations in text before tokenizing
ii) Similarly, It took words (name) as 3 different tokens and needed bracket removal before processing
iii) Example: Tree’s became 2 different tokens like tree and ‘s. Stemming or punctuation removal would have solved this issue.

Part2: Vector-space based IR System

I. **Query** = "Mountains and Railroad track to visit"

----- Results ------
1. Atkinson and Northern Railroad - 0.21829259240227494 - Relevant
2. Team track - 0.15012396424433647 - Relevant
3. Reading Blue Mountain and Northern Railroad - 0.1395739032716952 - Relevant
4. Saylyugem Mountains - 0.12850422741258666 - Relevant
5. Deseret Power Railroad - 0.1256664623018025 - Relevant
6. Vermont Railway - 0.12408688329027842 - Relevant
7. Dallas, Garland and Northeastern Railroad - 0.1067411607444787 - Relevant
8. Charles Paine - 0.10406985747465075 - Relevant
9. Emergency Broadcast System (album) - 0.10162511403700963 - Irrelevant
10. Samos, Lugo - 0.09800652588437607 – Relevant

II. **Query** = "All time favorite Music song tracks and duet ever recorded”

----- Results ------
1. Time Warp (album) - 0.1472503118789849 - Relevant
2. Tickson Music - 0.12998832342872196 - Relevant
3. Yes, I'm Ready - 0.12819683808775878 - Relevant
4. I Believe in You (Je crois en toi) - 0.12038268689330626- Relevant
5. Tout près du bonheur - 0.1193429742568333 - Relevant
6. Futari no Rocket - 0.11106431087226623 - Relevant
7. Would I Lie to You? (Eurythmics song) - 0.10428580806290534 - Relevant
8. Kaytanhousuja - 0.10420708806707804 - Relevant
9. Ever Blazin' - 0.10329924885927248 - Relevant
10. Illegal Alien (song) - 0.10279749429539703 – Relevant

III. **Query** = “President Bush of the United States of America”

----- Results ------
1. United States presidential election in Virginia, 2004 - 0.1857062648971375 - Relevant
2. United States presidential election in Maryland, 2004 - 0.18433102801588575 - Relevant
3. United States presidential election in Tennessee, 2004 - 0.18053794991144315 - Relevant
4. United States presidential election in Michigan, 2004 - 0.17952146026367582 - Relevant
5. Godot Waits For Homeland Security - 0.17199829647583945 - Irrelevant
6. Navigators (cycling team) - 0.1661245586244161 - Irrelevant

7. Douglas Little - 0.16414249488411503 - Irrelevant
8. Alma Adamkienė - 0.1567091796168298 - Irrelevant
9. Peter Keisler - 0.15422281166062637 - Irrelevant
10. Picture Rocks - 0.1510852236805083 - Irrelevant

IV. **Query** = "The carbon-fluorine bond is commonly found in pharmaceuticals and agrochemicals because it is generally metabolically stable"

----- Results ------
1. Organofluorine chemistry - 0.12480174802856096 - Relevant
2. Refeeding syndrome - 0.04872126560213592 - Irrelevant
3. Digenite - 0.04567740503630666 - Irrelevant
4. Cashiering - 0.04152624911039716 - Irrelevant
5. Ori (Hebrew) - 0.04101174302190606 - Irrelevant
6. Apis andreniformis - 0.03960919991162128 - Irrelevant
7. Flavones - 0.03954871452687718 - Irrelevant
8. Kitanofuji Katsuaki - 0.036813903568771705 - Irrelevant
9. Sex and Death 101 - 0.036662580095342276 - Irrelevant
10. Ulmus szechuanica - 0.03650604078548633 - Irrelevant

V. **Query** = "The Oddfest is an American celebration of comedy, music and artists that brings together some of today's funniest musical comedy acts"

----- Results ------
1. The Oddfest - 0.23497242828518217 - Relevant
2. George Dodd (Australian writer) - 0.10922926291600726 - Irrelevant
3. Chris Porter (comedian) - 0.10816979847934406 - Irrelevant
4. Edinburgh Churches Together - 0.08389019238952718 - Irrelevant
5. Metaphorical Music - 0.0828155988826628 - Irrelevant
6. Kabaret OT.TO - 0.08109457287140326 - Irrelevant
7. Henry Pottinger Stephens - 0.07758520900946762 - Irrelevant
8. Jason Stuart - 0.07607468945504206 - Irrelevant
9. Bob McClurg - 0.07470620263932548 - Irrelevant
10. Curtis Walker - 0.07281396431722455 - Irrelevant

VI. **Query** = "Henry Barber Richardson American archer won two Olympic bronze medals at two different editions of Olympic Games as well as youngest medallist at Summer Olympics at age of 15 years and 124 days. He also entered the Continental Style event, placing 15th with 171 points"

----- Results ------
1. Henry B. Richardson - 0.45297703207950313 - Relevant
2. Eugène Richez - 0.24967205589282598 - Irrelevant
3. Albert Dauchez - 0.219412469504806 - Irrelevant

4. Charles Quervel - 0.219412469504806 - Irrelevant
5. Louis-Albert Salingré - 0.21894216795297256 - Irrelevant
6. Charles Aubras - 0.21639546981583338 - Irrelevant
7. Eugène Grisot - 0.21297733647627606 - Irrelevant
8. Henri Berton - 0.2100007218194126 - Irrelevant
9. John Keyworth - 0.20919038670973172 - Irrelevant
10. Oscar Jay - 0.2079392486335594 – Irrelevant

VII. **Query** = "Golf Resort Tycoon business simulation computer game based premise players constructing own golf resorts limited amount funds Instant Action Challenges"

----- Results ------
1. Golf Resort Tycoon - 0.5166967675090071 - Relevant
2. Computer simulation and organizational studies - 0.09153813659852934 - Irrelevant
3. Virtual Pool 64 - 0.08825076876984311 - Irrelevant
4. Kenny Knox - 0.07715636733935873 - Irrelevant
5. J. L. Lewis - 0.06860060665210235 - Irrelevant
6. Battle Circuit - 0.0678278890320657 - Irrelevant
7. Pebble Beach Road Races - 0.06449178042377926- Irrelevant
8. Rik Massengale - 0.060265224615404525 - Irrelevant
9. Rugrats: Castle Capers - 0.05783751245982671 - Irrelevant
10. Evolution Snowboarding - 0.05570943112024132 – Irrelevant

VIII. **Query** = "Baltimore Jewish Times"

----- Results ------
1. Baltimore Jewish Times - 0.25872019518491507 - Relevant
2. The Detroit Jewish News - 0.22698564477833263 - Relevant
3. Knock three times - 0.1688230333980974 – Irrelevant
4. Baltimore Metros - 0.13219415465510506 - Irrelevant
5. Washington Metros - 0.10751444126423694 - Irrelevant
6. Lisa Suhay - 0.10146943556540011 – Irrelevant
7. David Ben Hassin - 0.0990473398350151 – Irrelevant
8. Belvedere Records - 0.09550365332982287 – Irrelevant
9. Moshe Sherer - 0.08723158446822835 – Irrelevant
10. Deuterocohnia - 0.08647798391207147 – Irrelevant

IX. **Query** = "Bressington"

----- Results ------
1. Alastair Bressington - 0.1663285880082713 - Relevant