**NLP Homework 1**

**Name - Karan Puran Ashar NetID – kashar**

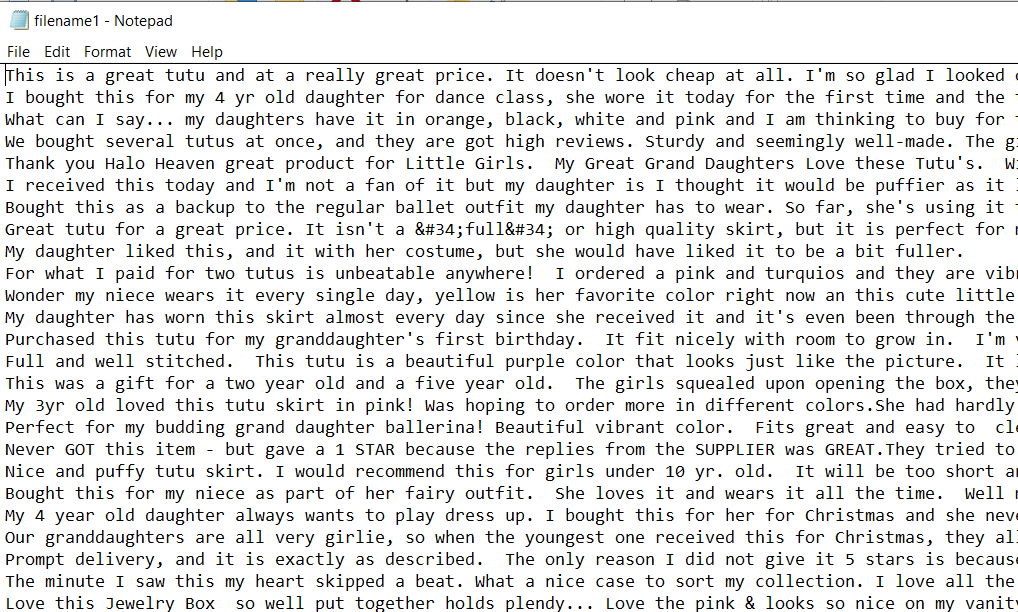
**Table of Contents –**1) Data Pre-processing  
2) Result of Analysis Task  
 2.1 Top 50 Words by Frequency  
 2.2 Top 50 Bigrams by Frequency  
 2.3 Top 50 Bigrams by PMI  
3) Interpretation Results   
4) Additional Analysis

**1)Data pre-processing**

I extracted the reviewText from the given file using regular expressions. Below attached is the screenshot of the code used to extract the text.



Below is the screenshot of the “filename1.txt” which was created.



I then created a corpus using PlaintextCorpusReader and tokenized the corpus using nltk.word\_tokenize.

For the first task I decided to use the isalpha() alpha function to filter the data first. I then removed stop words using the built in stopwords list. I wanted to keep all upper-case words because I think to write reviews people use upper-case to emphasize on what they are saying. I did not want to lose that aspect. To remove all capital stopwords I created a new list stopwordsUpper using the capitalize function. I then removed the capital stopwords using this list.

For the second and third task I first created the bigrams using BigramCollocationFinder. I then applied the Alpha filter which is filtering out the non-alpha words. I then filtered out words using stopwords and stopwordsUpper.

For the third task I also applied a frequency filter set to 5.

**2) Results from Analysis task**

2.1 - Top 50 words by frequency

('like', 87738)

('size', 83282)

('wear', 78845)

('fit', 77125)

('great', 65445)

('would', 63288)

('comfortable', 55353)

('well', 54539)

('good', 54105)

('one', 52962)

('shoes', 49574)

('love', 47390)

('really', 43427)

('little', 42321)

('look', 41726)

('nice', 40757)

('bought', 38196)

('get', 37906)

('price', 33953)

('pair', 33365)

('color', 32649)

('time', 32330)

('small', 31923)

('quality', 31859)

('ordered', 31306)

('shoe', 30752)

('much', 28955)

('perfect', 28385)

('buy', 27843)

('made', 27290)

('got', 26588)

('watch', 26088)

('bit', 25803)

('long', 25005)

('feet', 24897)

('looks', 24752)

('also', 24245)

('even', 23492)

('fits', 22847)

('recommend', 22556)

('back', 22520)

('work', 22236)

('still', 22029)

('wearing', 21938)

('big', 21937)

('looking', 21017)

('could', 20797)

('pretty', 20450)

('right', 20337)

('material', 20204)



2.2 - Top 50 bigrams by frequency

(('well', 'made'), 0.0004223229196771571)

(('would', 'recommend'), 0.00027974470381769305)

(('good', 'quality'), 0.00025981497510449056)

(('really', 'like'), 0.00025138516818141277)

(('fit', 'perfectly'), 0.00022286952500951996)

(('look', 'like'), 0.00021938312338083964)

(('fit', 'well'), 0.0002168854028110388)

(('highly', 'recommend'), 0.00020783116574551078)

(('another', 'pair'), 0.0001980484268471242)

(('look', 'great'), 0.00018743311442547068)

(('looks', 'great'), 0.0001821254582146439)

(('year', 'old'), 0.00018129288469137696)

(('feel', 'like'), 0.00017962773764484307)

(('looks', 'like'), 0.00016937667613961882)

(('usually', 'wear'), 0.0001590735787891904)

(('fit', 'great'), 0.00015537903377969335)

(('normally', 'wear'), 0.00014648090424977789)

(('long', 'time'), 0.00014356689691834358)

(('light', 'weight'), 0.0001402366028252758)

(('one', 'size'), 0.0001382592407075168)

(('every', 'day'), 0.00013789498979108753)

(('even', 'though'), 0.00013659409366098294)

(('arch', 'support'), 0.00013617780689934947)

(('size', 'larger'), 0.00013362805048434444)

(('look', 'good'), 0.00013060997146250176)

(('little', 'bit'), 0.00012837243011872186)

(('half', 'size'), 0.00012472992095442897)

(('great', 'price'), 0.00012348106066952856)

(('first', 'time'), 0.00012030687411207334)

(('fits', 'well'), 0.00011942226474360221)

(('would', 'buy'), 0.00011879783460115202)

(('much', 'better'), 0.00011375035761634616)

(('fits', 'perfectly'), 0.00011338610669991688)

(('really', 'nice'), 0.00011224131810542483)

(('different', 'colors'), 0.00011182503134379135)

(('long', 'enough'), 0.00010776623541786501)

(('high', 'quality'), 0.00010708976943021062)

(('flip', 'flops'), 0.00010464408470561397)

(('looks', 'good'), 0.00010152193399336294)

(('second', 'pair'), 9.886810588794955e-05)

(('size', 'smaller'), 9.444505904559392e-05)

(('perfect', 'fit'), 9.173919509497634e-05)

(('another', 'one'), 9.158308755936379e-05)

(('really', 'cute'), 9.12188366429345e-05)

(('buy', 'another'), 9.101069326211777e-05)

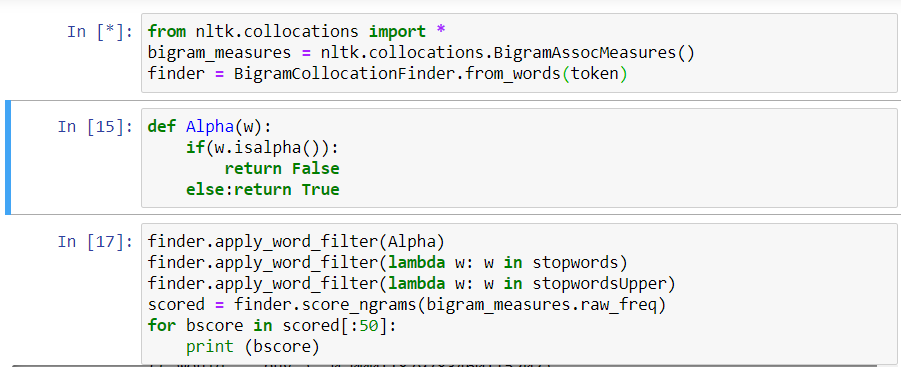
(('first', 'pair'), 9.002201220323827e-05)

(('fits', 'great'), 9.002201220323827e-05)

(('many', 'compliments'), 8.89292594539504e-05)

(('years', 'ago'), 8.840890100190856e-05)

(('fit', 'perfect'), 8.747225578823326e-05)



2.3 -Top 50 bigrams by PMI and frequency filter set to 5

(('ABCO', 'TECH'), 24.195918984257116)

(('ABD', 'SWIM'), 24.195918984257116)

(('AEGIS', 'microbe'), 24.195918984257116)

(('ANDS', 'COMFORATABLE'), 24.195918984257116)

(('ANN', 'KLIEN'), 24.195918984257116)

(('ANNE', 'MARIE'), 24.195918984257116)

(('ARIAT', 'ROPER'), 24.195918984257116)

(('ARRAGE', 'REMOVING'), 24.195918984257116)

(('ATTHIS', 'TIMEI'), 24.195918984257116)

(('AVID', 'PITTSBURGH'), 24.195918984257116)

(('Abstract', 'Algebra'), 24.195918984257116)

(('Addis', 'Ababa'), 24.195918984257116)

(('Adina', 'Expanding'), 24.195918984257116)

(('AdjustableBreathable', 'MaterialOne'), 24.195918984257116)

(('Aer', 'Lingus'), 24.195918984257116)

(('AfterglowFlash', 'AlertFlashes'), 24.195918984257116)

(('Alien', 'Spaceship'), 24.195918984257116)

(('Allout', 'Blaze'), 24.195918984257116)

(('Angie', 'Harmon'), 24.195918984257116)

(('AntimicrobialMoisture', 'resistantClose'), 24.195918984257116)

(('Approach', 'otlino'), 24.195918984257116)

(('Archduke', 'Franz'), 24.195918984257116)

(('Ashtanga', 'vinyasa'), 24.195918984257116)

(('AssnCalvin', 'KleinThe'), 24.195918984257116)

(('Asuka', 'Langley'), 24.195918984257116)

(('Avanti', 'Superchannels'), 24.195918984257116)

(('Awning', 'StripeThis'), 24.195918984257116)

(('BACOVA', 'GUILD'), 24.195918984257116)

(('BALES', 'RUBBING'), 24.195918984257116)

(('BATTLE', 'SITUATIONS'), 24.195918984257116)

(('BEEFY', 'TSHIT'), 24.195918984257116)

(('BEN', 'DAVIS'), 24.195918984257116)

(('BNS', 'TARDES'), 24.195918984257116)

(('BORAT', 'MANKINI'), 24.195918984257116)

(('BUG', 'REPELLANT'), 24.195918984257116)

(('BUL', 'OUS'), 24.195918984257116)

(('BUN', 'HAIRSTYLE'), 24.195918984257116)

(('BURNT', 'SIENNA'), 24.195918984257116)

(('BUTG', 'UNFORUNTLY'), 24.195918984257116)

(('BabyLuvable', 'FriendsBrands'), 24.195918984257116)

(('BabyRalph', 'LaurenUS'), 24.195918984257116)

(('BagGood', 'PriceSuper'), 24.195918984257116)

(('BandLED', 'backlightSelectable'), 24.195918984257116)

(('Bangggg', 'Bleeep'), 24.195918984257116)

(('Battlestar', 'Galactica'), 24.195918984257116)

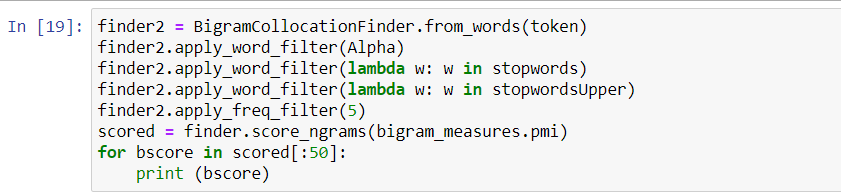
(('Beaba', 'Stackable'), 24.195918984257116)

(('Beaux', 'Bijoux'), 24.195918984257116)

(('Bein', 'Vampy'), 24.195918984257116)

(('Bernie', 'Mev'), 24.195918984257116)

(('Bibidi', 'Bobidi'), 24.195918984257116)



**3) Interpretation results**

First let’s look at the word frequency output. We can see a lot of positive words like ‘like’, ‘great’, ‘comfortable’, ‘good’, ’love’, ‘nice’, ‘perfect’, ‘recommend’, ‘pretty’. This seems to indicate that majority of the reviews are positive. People have liked what they bought and are recommending the product to other people. But this is only an assumption we can make. We cannot conclude our statement by only seeing the word frequency.  
There are many words which are related to something you wear. For example, words like ‘size’, ‘wear’, ‘fit’, ‘shoe’, ‘feet’, ‘material’ indicate that the reviews are about clothing and apparels.  
The word ‘price’ is also seen many times. But we do not know in what context is it being used. For example, high price or low price.

Now let us look at the most frequent bigrams we created. As we can see we get a clearer picture of the text. Words like ‘recommend’, ’quality’, ‘great’, ‘perfect’ and preceded by ‘fit’, ’looks’, ’highly’, ‘good’ which indicates the people are happy with the products.   
I also noticed the word ‘size’ occurring may times. We cannot yet conclude to what the people are exactly trying to say about the size. For example, bigrams like ‘size smaller’, ‘size larger’, ’one size’, ’half size’ do not give us a picture ‘as a whole’. We can say there seems to be some issue regarding the size with a few people. But to some the size fits perfectly.

The bigrams created in both the tasks have completely different outputs. The first task gives us the frequency and the second gives us the PMI score. The bigram of the second task probably gives us names of the company to which the products belong.

**4) Additional Analysis**

I can focus on the word ‘size’ and get to know what the users are trying to say. As we can see there are mixed reviews about the size, so I can try to figure out the products for the reviews and see which specific product has size issue and which product is a perfect fit. We can also use stemming to merge words like ‘looks’ to ‘look’ to get better results.