Amazon Fine Food Reviews Analysis Truncated SVD

In [0]:
```
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")


import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer

import re
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle
```

```
from tqdm import tqdm
import os
```

In [5]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?
client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleuser
content.com&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&scope=emai
l%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdocs.test%20https%3A%2F%2
Fwww.googleapis.com%2Fauth%2Fdrive%20https%3A%2F%2Fwww.googleapis.com%2
Fauth%2Fdrive.photos.readonly%20https%3A%2F%2Fwww.googleapis.com%2Faut
h%2Fpeopleapi.readonly&response_type=code

Enter your authorization code:
··········
Mounted at /content/drive
```

In [0]:
```
!cp "/content/drive/My Drive/final.sqlite" "final.sqlite"
```

In [7]:
```
import os
if os.path.isfile('final.sqlite'):
    conn = sqlite3.connect('final.sqlite')
    final = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score !=
 3 """, conn)
    conn.close()
else:
    print("Please the above cell")

print("Preprocessed Amzon fine food data columns  shape : ",final.shape
)
print("fPreprocessed Amzon fine food data columns       :",final.column
s.values)
```

```
Preprocessed Amzon fine food data columns  shape :  (364171, 12)
fPreprocessed Amzon fine food data columns       : ['index' 'Id' 'Produ
ctId' 'UserId' 'ProfileName' 'HelpfulnessNumerator'
```

```
 'HelpfulnessDenominator' 'Score' 'Time' 'Summary' 'Text' 'CleanedTex
t']
```

In [8]:
```
preprocessed=final['CleanedText'][:100000]
len(preprocessed)
```

Out[8]:
```
100000
```

TFIDF Vectorization

In [9]:
```
tf_idf_vect = TfidfVectorizer(min_df=10,max_df=0.9,max_features=2000)
tf_idf_vect.fit(preprocessed)
print("some sample features(unique words in the corpus)",tf_idf_vect.ge
t_feature_names()[0:10])
print('='*50)

final_tf_idf = tf_idf_vect.transform(preprocessed)
print("the type of count vectorizer ",type(final_tf_idf))
print("the shape of out text TFIDF vectorizer ",final_tf_idf.get_shape
())
print("the number of unique words including both unigrams and bigrams "
, final_tf_idf.get_shape()[1])
```

```
some sample features(unique words in the corpus) ['abl', 'about', 'abo
v', 'absolut', 'absorb', 'accept', 'accord', 'acid', 'acquir', 'acros
s']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer  (100000, 2000)
the number of unique words including both unigrams and bigrams  2000
```

Take top 2000 or 3000 features from tf-idf vectorizers using idf_ score.

In [10]:
```
indices = np.argsort(tf_idf_vect.idf_)[::-1]
```

```
features = tf_idf_vect.get_feature_names()

top_features = [features[i] for i in indices[:3000]]
print(top_features)
```

```
['tunnel', 'matcha', 'anchovi', 'lobster', 'gopher', 'flea', 'goji', 'p
ellet', 'evo', 'darjeel', 'film', 'xylitol', 'chipotl', 'couscous', 'ae
rogarden', 'fenugreek', 'bulli', 'lavend', 'clam', 'bpa', 'rooibo', 'go
at', 'coke', 'jug', 'pine', 'refri', 'purina', 'pear', 'ketchup', 'rin
g', 'cough', 'seawe', 'canida', 'dye', 'felin', 'bergamot', 'muesli',
'pearl', 'everlast', 'tablet', 'infect', 'rabbit', 'rum', 'oreo', 'hem
p', 'wasabi', 'cancer', 'cola', 'pineappl', 'eden', 'balsam', 'moth',
'princ', 'king', 'mate', 'cooker', 'mrs', 'cashew', 'malt', 'jalapeno',
'cramp', 'tinkyada', 'pig', 'greeni', 'numi', 'spearmint', 'quart', 'de
caffein', 'leg', 'british', 'apricot', 'jim', 'larabar', 'greek', 'roya
l', 'fudg', 'valley', 'kashi', 'pastri', 'japan', 'hip', 'hotter', 'ran
cid', 'miso', 'catnip', 'leak', 'death', 'european', 'stain', 'dirt',
'win', 'alcohol', 'hummus', 'flaxse', 'zuke', 'infus', 'newman', 'bell
i', 'calcium', 'passion', 'tray', 'oolong', 'corner', 'recipi', 'dispen
s', 'net', 'workout', 'fatti', 'camp', 'poop', 'mole', 'parmesan', 'co
w', 'fishi', 'starch', 'master', 'virgin', 'york', 'pad', 'shed', 'weig
h', 'tsp', 'drip', 'hodgson', 'exercis', 'smart', 'casserol', 'east',
'kona', 'skip', 'poison', 'soil', 'crap', 'twin', 'ramen', 'mislead',
'build', 'emerg', 'minor', 'durabl', 'dream', 'smokey', 'harm', 'exclu
s', 'bay', 'giant', 'occupi', 'simmer', 'bet', 'zico', 'nurs', 'sweetne
r', 'student', 'kong', 'spinach', 'canist', 'tub', 'florida', 'intact',
'darn', 'feet', 'seafood', 'satur', 'symptom', 'wood', 'scare', 'oppo
s', 'rubber', 'cancel', 'materi', 'ratio', 'promot', 'molass', 'readil
i', 'fond', 'joy', 'featur', 'puff', 'soooo', 'kidney', 'blender', 'nyl
abon', 'page', 'batter', 'certifi', 'dash', 'club', 'graham', 'pickl',
'trade', 'unusu', 'cap', 'ham', 'studi', 'collect', 'germani', 'itali',
'beg', 'cupcak', 'terrier', 'mold', 'tabasco', 'hershey', 'vacat', 'upd
at', 'appli', 'father', 'fight', 'twist', 'thicken', 'refil', 'shini',
'america', 'burnt', 'happili', 'island', 'gritti', 'europ', 'stumbl',
'acquir', 'chamomil', 'hunt', 'biggest', 'bonsai', 'smoki', 'ultim', 'a
sid', 'dental', 'nectar', 'frustrat', 'weekend', 'up', 'insan', 'hydrog
en', 'rope', 'guilti', 'advic', 'metal', 'plump', 'graini', 'photo', 's
tove', 'refer', 'key', 'hey', 'sooo', 'numer', 'fault', 'waffl', 'fir
e', 'india', 'carton', 'fool', 'bran', 'sunflow', 'mocha', 'brush', 're
scu', 'shouldnt', 'sun', 'jack', 'creme', 'dehydr', 'chose', 'reseal',
```

```
'hell', 'ami', 'citrus', 'refin', 'behind', 'clove', 'breed', 'mash',
'bacon', 'burger', 'somehow', 'sampler', 'stool', 'wholesom', 'habit',
'indulg', 'factor', 'rye', 'scienc', 'reciev', 'absorb', 'kitten', 'bas
il', 'tortilla', 'barri', 'highest', 'carbohydr', 'googl', 'minim', 'su
shi', 'beer', 'below', 'pair', 'squeez', 'danger', 'consider', 'cure',
'tran', 'chain', 'lamb', 'regard', 'rough', 'wafer', 'till', 'generou
s', 'theyd', 'theyll', 'unsweeten', 'paw', 'mile', 'transit', 'sore',
'macaroni', 'slim', 'barley', 'request', 'colleg', 'neighbor', 'wrote',
'familiar', 'marshmallow', 'bug', 'advis', 'lentil', 'san', 'harvest',
'boyfriend', 'appetit', 'superb', 'shown', 'challeng', 'thorough', 'pre
ssur', 'creamer', 'suck', 'fashion', 'cupboard', 'tummi', 'kettl', 'dow
nsid', 'partial', 'miner', 'hadnt', 'act', 'desk', 'diarrhea', 'chef',
'supplier', 'recal', 'toddler', 'shame', 'hamburg', 'track', 'yeah', 'y
ard', 'earlier', 'answer', 'bakeri', 'cafe', 'ladi', 'walnut', 'rub',
'altoid', 'texa', 'space', 'decor', 'info', 'class', 'punch', 'irish',
'kinda', 'spit', 'sudden', 'weather', 'wrapper', 'grade', 'lead', 'yest
erday', 'fluffi', 'unpleas', 'storag', 'vomit', 'latt', 'link', 'choles
terol', 'jasmin', 'push', 'garbag', 'dice', 'closest', 'target', 'koshe
r', 'antioxid', 'fake', 'goodi', 'squar', 'buyer', 'titl', 'manner', 'g
entl', 'grape', 'meant', 'risk', 'lick', 'elimin', 'foil', 'fail', 'no
n', 'explain', 'excess', 'pecan', 'ridicul', 'rose', 'stori', 'counte
r', 'ideal', 'speak', 'tazo', 'mellow', 'degre', 'affect', 'savori', 'e
ar', 'faster', 'compliment', 'suspect', 'frank', 'stress', 'loss', 'pen
ni', 'iron', 'entertain', 'overnight', 'tasteless', 'purs', 'peel', 'th
ru', 'munch', 'eventu', 'ador', 'automat', 'pink', 'moistur', 'stuf',
'lollipop', 'meet', 'match', 'kernel', 'mountain', 'magic', 'gobbl', 'f
ragrant', 'solut', 'dad', 'quinoa', 'jump', 'plate', 'moder', 'cane',
'household', 'dump', 'truffl', 'pinch', 'lucki', 'afraid', 'wed', 'stre
ngth', 'filter', 'broke', 'wake', 'inde', 'tofu', 'wateri', 'protect',
'thicker', 'forc', 'exampl', 'saver', 'teabag', 'fanci', 'lighter', 'te
ar', 'section', 'brother', 'tight', 'buffalo', 'rang', 'shrimp', 'secre
t', 'prior', 'clump', 'edg', 'decad', 'repeat', 'somewher', 'butteri',
'god', 'rise', 'eas', 'report', 'marinad', 'support', 'aromat', 'econo
m', 'split', 'attract', 'strip', 'adjust', 'winner', 'odor', 'spring',
'crystal', 'zero', 'fructos', 'anytim', 'citi', 'grocer', 'specialti',
'fli', 'oili', 'shred', 'destroy', 'reward', 'further', 'necessari', 'r
awhid', 'moment', 'maintain', 'cute', 'commerci', 'program', 'hurt', 'e
lsewher', 'gain', 'dust', 'trash', 'pup', 'robust', 'gallon', 'pleasu
r', 'rip', 'sharp', 'hazelnut', 'mark', 'min', 'swallow', 'essenti', 'j
```

```
am', 'steam', 'intend', 'fulli', 'childhood', 'spoil', 'throughout', 'f
lax', 'softer', 'appeal', 'tangi', 'rid', 'inexpens', 'taco', 'grass',
'fortun', 'thus', 'bright', 'snap', 'whip', 'breast', 'vari', 'beyond',
'nutrient', 'lol', 'therefor', 'bbq', 'german', 'step', 'settl', 'catc
h', 'bubbl', 'accept', 'choke', 'sugari', 'ruin', 'messi', 'indic', 'gr
easi', 'refus', 'layer', 'temperatur', 'edibl', 'regret', 'against', 'v
ersatil', 'south', 'liver', 'cranberri', 'england', 'calm', 'discount',
'inch', 'press', 'dead', 'slow', 'costco', 'bore', 'japanes', 'mug', 's
lowli', 'sweeter', 'stew', 'effort', 'tongu', 'occas', 'fed', 'filler',
'wide', 'experienc', 'hasnt', 'attent', 'iti', 'anni', 'flat', 'boost',
'bewar', 'funni', 'term', 'mango', 'reorder', 'pump', 'dens', 'skepti
c', 'promis', 'harder', 'neither', 'method', 'attempt', 'litter', 'into
ler', 'owner', 'question', 'rais', 'man', 'mushi', 'parent', 'reactio
n', 'pud', 'rins', 'shelv', 'complex', 'trader', 'salsa', 'cent', 'hid
e', 'diagnos', 'theyv', 'bargain', 'grate', 'member', 'among', 'drain',
'chewer', 'via', 'bold', 'center', 'ahead', 'shock', 'signific', 'hol
e', 'flavour', 'unabl', 'dough', 'locat', 'leftov', 'disappear', 'heart
i', 'sausag', 'spray', 'forev', 'toler', 'medic', 'bonus', 'desir', 'cr
ack', 'multipl', 'california', 'bigelow', 'fruiti', 'tip', 'reach', 'me
mori', 'onto', 'sprout', 'frost', 'assort', 'awhil', 'palat', 'msg', 'e
mail', 'nose', 'agav', 'discontinu', 'handi', 'sardin', 'gravi', 'cub
e', 'floor', 'mushroom', 'defin', 'kitti', 'kraft', 'intak', 'smoothi',
'carrot', 'mexican', 'caught', 'movi', 'rock', 'drive', 'pouch', 'was
h', 'human', 'unit', 'hear', 'comparison', 'root', 'common', 'web', 'ba
si', 'pamela', 'activ', 'kill', 'sesam', 'splenda', 'crush', 'suffer',
'throat', 'sign', 'sens', 'measur', 'besid', 'design', 'celesti', 'ever
ywher', 'awar', 'front', 'usa', 'hungri', 'senseo', 'kibbl', 'forget',
'purpos', 'disgust', 'finger', 'tender', 'yogi', 'walmart', 'car', 'and
or', 'classic', 'lime', 'mini', 'grind', 'kit', 'concentr', 'young', 'l
id', 'dozen', 'assum', 'youd', 'odd', 'retail', 'peach', 'indian', 'ave
rag', 'mixtur', 'enhanc', 'gross', 'trick', 'crumb', 'frequent', 'gas',
'scent', 'scoop', 'practic', 'despit', 'specif', 'sip', 'prevent', 'buc
k', 'basket', 'respons', 'jelli', 'ten', 'began', 'town', 'mapl', 'gard
en', 'werent', 'gold', 'firm', 'comfort', 'tooth', 'joe', 'peppermint',
'medicin', 'tough', 'sister', 'girl', 'initi', 'mail', 'constant', 'imp
oss', 'middl', 'hair', 'hesit', 'wine', 'prime', 'lab', 'develop', 'old
er', 'accord', 'sooth', 'junk', 'pea', 'shot', 'outstand', 'manag', 'sl
eep', 'bed', 'trust', 'luck', 'ran', 'lbs', 'bill', 'die', 'breath', 'y
east', 'refund', 'flake', 'caramel', 'folk', 'doubt', 'loaf', 'diseas',
```

```
'strang', 'popular', 'asian', 'overwhelm', 'haribo', 'period', 'subscri
pt', 'post', 'china', 'drank', 'bare', 'distinct', 'curri', 'tuna', 'ra
re', 'pocket', 'guest', 'face', 'broth', 'steak', 'mac', 'authent', 'ra
ve', 'terrif', 'choos', 'earli', 'golden', 'spaghetti', 'spent', 'corre
ct', 'success', 'turkey', 'fell', 'deep', 'wors', 'vendor', 'mistak',
'grown', 'empti', 'complain', 'doctor', 'earth', 'appar', 'nutriti', 'd
issolv', 'twine', 'load', 'cardboard', 'superior', 'walk', 'pork', 'int
ens', 'taken', 'beverag', 'chines', 'wet', 'crust', 'spot', 'supplemen
t', 'dent', 'stevia', 'oat', 'alot', 'weak', 'shell', 'cheddar', 'nor',
'bud', 'espresso', 'dairi', 'heart', 'air', 'liter', 'premium', 'pumpki
n', 'soak', 'bother', 'damag', 'sea', 'favor', 'sad', 'lipton', 'starbu
ck', 'nasti', 'pizza', 'ton', 'upset', 'posit', 'weird', 'third', 'yell
ow', 'separ', 'raisin', 'mustard', 'freezer', 'eye', 'wild', 'proper',
'begin', 'talk', 'book', 'equal', 'internet', 'relax', 'oven', 'brown
i', 'major', 'stronger', 'farm', 'select', 'tree', 'limit', 'anim', 'ro
und', 'contact', 'pain', 'threw', 'grab', 'stash', 'admit', 'adult', 'p
ill', 'afford', 'themselv', 'soda', 'biscuit', 'bigger', 'pull', 'flowe
r', 'grill', 'becam', 'solid', 'crispi', 'claim', 'whenev', 'head', 'ta
bl', 'negat', 'holiday', 'shes', 'earl', 'delic', 'toss', 'shake', 'nut
ti', 'occasion', 'guy', 'depend', 'thai', 'scratch', 'tire', 'leaf', 'o
kay', 'handl', 'teaspoon', 'afternoon', 'abov', 'school', 'paper', 'wor
st', 'countri', 'winter', 'tart', 'fabul', 'sticki', 'bunch', 'chunk',
'child', 'smoke', 'sorri', 'heaven', 'eater', 'pretzel', 'gram', 'subt
l', 'current', 'whether', 'blueberri', 'introduc', 'chop', 'obvious',
'burn', 'upon', 'birthday', 'muffin', 'warn', 'remain', 'everyday', 'sy
stem', 'tin', 'vitamin', 'inform', 'boy', 'advertis', 'salmon', 'raspbe
rri', 'freez', 'granola', 'banana', 'frozen', 'pass', 'grey', 'celiac',
'allerg', 'honest', 'medium', 'charg', 'door', 'portion', 'fix', 'powe
r', 'known', 'english', 'stuck', 'relat', 'decaf', 'chanc', 'trap', 'it
alian', 'yum', 'pricey', 'reduc', 'vinegar', 'blood', 'fridg', 'gotte
n', 'offic', 'lost', 'felt', 'spread', 'poor', 'futur', 'count', 'appea
r', 'extract', 'stapl', 'age', 'lack', 'apart', 'safe', 'various', 'cre
at', 'descript', 'style', 'gourmet', 'spoon', 'overpow', 'late', 'somew
hat', 'write', 'sampl', 'hes', 'understand', 'room', 'troubl', 'spend',
'plenti', 'pancak', 'travel', 'dessert', 'berri', 'american', 'crumbl',
'tablespoon', 'parti', 'present', 'wow', 'improv', 'prompt', 'herbal',
'forward', 'decent', 'dollar', 'standard', 'vegetarian', 'lose', 'strai
ght', 'sound', 'maker', 'refresh', 'job', 'drinker', 'paid', 'pod', 'su
mmer', 'crisp', 'yogurt', 'ball', 'grew', 'sensit', 'higher', 'vegan',
```

```
'strawberri', 'aw', 'youv', 'steep', 'pan', 'plant', 'heard', 'word',
'comment', 'heavi', 'alon', 'pie', 'acid', 'diabet', 'site', 'skin', 'h
orribl', 'thrill', 'mess', 'across', 'dress', 'typic', 'brought', 'othe
rwis', 'uniqu', 'test', 'incred', 'bob', 'children', 'knew', 'pantri',
'expir', 'visit', 'control', 'websit', 'perhap', 'outsid', 'remov', 're
friger', 'increas', 'crazi', 'doubl', 'bear', 'research', 'shelf', 'pla
y', 'concern', 'chai', 'form', 'instruct', 'digest', 'kept', 'appreci',
'fall', 'jerki', 'french', 'remind', 'complaint', 'sick', 'trip', 'manu
factur', 'aftertast', 'formula', 'kick', 'serious', 'stand', 'basic',
'imagin', 'sprinkl', 'chemic', 'preserv', 'blue', 'number', 'quantiti',
'cocoa', 'stir', 'stale', 'terribl', 'bland', 'describ', 'mint', 'raw',
'tend', 'level', 'gum', 'cool', 'six', 'whatev', 'pictur', 'readi', 'he
rb', 'learn', 'fri', 'main', 'train', 'unlik', 'puppi', 'avoid', 'liqui
d', 'toast', 'previous', 'short', 'shipment', 'lover', 'sour', 'botto
m', 'licoric', 'touch', 'hint', 'sort', 'tini', 'loos', 'mom', 'agre',
'seal', 'seller', 'hook', 'realiz', 'mother', 'fit', 'move', 'requir',
'allow', 'crunch', 'balanc', 'tradit', 'broken', 'bone', 'energi', 'ove
ral', 'lower', 'sale', 'dip', 'homemad', 'onion', 'glass', 'produc', 'u
nless', 'hate', 'opinion', 'mill', 'send', 'almond', 'veggi', 'impres
s', 'picki', 'none', 'oatmeal', 'immedi', 'grow', 'matter', 'machin',
'sandwich', 'condit', 'beauti', 'healthier', 'anymor', 'slice', 'anywa
y', 'vet', 'suppli', 'possibl', 'yourself', 'amazoncom', 'shape', 'twic
e', 'fun', 'kitchen', 'consum', 'weve', 'roll', 'sourc', 'ounc', 'cream
i', 'warm', 'wrap', 'anywher', 'busi', 'under', 'batch', 'plan', 'garli
c', 'cherri', 'orang', 'crave', 'thin', 'general', 'moist', 'boil', 'sh
ow', 'appl', 'gummi', 'caffein', 'juic', 'pour', 'switch', 'chili', 'to
day', 'five', 'excit', 'cheap', 'allergi', 'import', 'happen', 'cover',
'true', 'clear', 'pure', 'seen', 'option', 'popcorn', 'larger', 'instan
t', 'easier', 'do', 'truli', 'drop', 'provid', 'carb', 'mind', 'note',
'told', 'simpl', 'interest', 'fiber', 'sodium', 'beat', 'ginger', 'bod
i', 'melt', 'unfortun', 'daili', 'worri', 'idea', 'return', 'restaur',
'stomach', 'subscrib', 'togeth', 'arent', 'microwav', 'aroma', 'figur',
'delight', 'fish', 'singl', 'mention', 'thick', 'point', 'tomato', 'ben
efit', 'difficult', 'fair', 'artifici', 'toy', 'coat', 'suppos', 'worl
d', 'hold', 'oliv', 'sold', 'hit', 'caus', 'given', 'ground', 'dinner',
'pleasant', 'teeth', 'watch', 'chewi', 'name', 'yes', 'due', 'process',
'miss', 'veget', 'rest', 'pot', 'pop', 'share', 'sent', 'content', 'lat
er', 'individu', 'combin', 'babi', 'pet', 'ate', 'roast', 'soon', 'bow
l', 'itself', 'done', 'christma', 'lemon', 'label', 'rate', 'four', 'st
```

```
ate', 'break', 'supermarket', 'beef', 'lunch', 'area', 'mine', 'entir',
'noodl', 'certain', 'deliveri', 'line', 'saw', 'shop', 'within', 'insi
d', 'gone', 'finish', 'custom', 'between', 'effect', 'along', 'date',
'cinnamon', 'deliv', 'wrong', 'except', 'discov', 'wast', 'grain', 'sub
stitut', 'huge', 'smaller', 'throw', 'mild', 'plastic', 'eaten', 'someo
n', 'egg', 'left', 'follow', 'bring', 'nutrit', 'similar', 'consid', 'p
acket', 'soy', 'clean', 'servic', 'sit', 'alreadi', 'weight', 'extrem',
'particular', 'issu', 'check', 'rememb', 'flour', 'wife', 'replac', 'se
ed', 'guess', 'special', 'fantast', 'addict', 'brew', 'suggest', 'dur
e', 'pick', 'bulk', 'simpli', 'potato', 'syrup', 'set', 'search', 'beco
m', 'total', 'final', 'cours', 'salti', 'prepar', 'result', 'nut', 'coc
onut', 'plain', 'awesom', 'choic', 'pepper', 'continu', 'version', 'mea
n', 'super', 'pound', 'bite', 'night', 'daughter', 'wouldnt', 'wait',
'stay', 'feed', 'conveni', 'easili', 'jar', 'crunchi', 'brown', 'ask',
'complet', 'valu', 'normal', 'near', 'altern', 'cannot', 'consist', 'of
fer', 'direct', 'wasnt', 'sweeten', 'base', 'yummi', 'rather', 'close',
'peanut', 'cake', 'honey', 'salad', 'smooth', 'experi', 'fan', 'often',
'list', 'goe', 'stock', 'origin', 'cracker', 'cold', 'cut', 'exact', 'c
orn', 'life', 'pay', 'color', 'havent', 'dark', 'etc', 'youll', 'heat',
'soft', 'bitter', 'son', 'dish', 'side', 'person', 'addit', 'spici', 'p
art', 'cereal', 'believ', 'mouth', 'recent', 'stick', 'rich', 'myself',
'past', 'pasta', 'onlin', 'plus', 'meat', 'sell', 'includ', 'everyon',
'slight', 'although', 'sometim', 'breakfast', 'notic', 'took', 'cream',
'hour', 'mayb', 'powder', 'cheaper', 'wheat', 'yet', 'either', 'run',
'tell', 'vanilla', 'els', 'hous', 'satisfi', 'couldnt', 'chang', 'blen
d', 'protein', 'second', 'must', 'gave', 'isnt', 'fine', 'type', 'gla
d', 'white', 'everyth', 'red', 'market', 'three', 'anyon', 'surpris',
'turn', 'fruit', 'extra', 'fact', 'season', 'ice', 'deal', 'fast', 'sto
p', 'wish', 'fat', 'next', 'call', 'longer', 'end', 'compar', 'went',
'care', 'spice', 'amaz', 'hope', 'chew', 'least', 'fill', 'might', 'mor
n', 'coupl', 'black', 'cost', 'leav', 'piec', 'ill', 'noth', 'health',
'larg', 'decid', 'him', 'half', 'chip', 'varieti', 'place', 'instead',
'prefer', 'light', 'such', 'money', 'whi', 'full', 'bottl', 'butter',
'theyr', 'ago', 'calori', 'hand', 'probabl', 'bean', 'gift', 'star', 'm
eal', 'let', 'live', 'chees', 'pleas', 'bread', 'especi', 'carri', 'pe
r', 'expect', 'soup', 'husband', 'said', 'cat', 'read', 'until', 'bak
e', 'avail', 'diet', 'gluten', 'kind', 'reason', 'absolut', 'wont', 'ab
l', 'save', 'salt', 'green', 'top', 'expens', 'his', 'minut', 'compan
i', 'bar', 'chicken', 'strong', 'recip', 'home', 'where', 'low', 'amoun
```

```
t', 'onc', 'own', 'away', 'disappoint', 'came', 'may', 'kid', 'serv',
'oil', 'receiv', 'usual', 'real', 'new', 'should', 'cooki', 'candi', 's
mell', 'open', 'around', 'pretti', 'go', 'week', 'rice', 'anyth', 'eac
h', 'both', 'item', 'size', 'worth', 'arriv', 'peopl', 'big', 'down',
'case', 'thank', 'thought', 'almost', 'sever', 'contain', 'friend', 'ba
d', 'quit', 'textur', 'famili', 'through', 'be', 'organ', 'dri', 'anot
h', 'problem', 'happi', 'howev', 'feel', 'sauc', 'help', 'start', 'regu
lar', 'actual', 'though', 'natur', 'definit', 'her', 'off', 'those', 'a
d', 'see', 'excel', 'far', 'groceri', 'old', 'tasti', 'healthi', 'mont
h', 'sure', 'ingredi', 'milk', 'quick', 'less', 'didnt', 'small', 'does
nt', 'whole', 'same', 'cook', 'enough', 'long', 'here', 'qualiti', 'see
m', 'she', 'cup', 'someth', 'while', 'back', 'right', 'last', 'snack',
'stuff', 'pack', 'easi', 'hard', 'water', 'hot', 'doe', 'chocol', 'di
d', 'without', 'fresh', 'coffe', 'free', 'put', 'ship', 'into', 'how',
'everi', 'perfect', 'take', 'alway', 'ever', 'few', 'got', 'befor', 'wo
nder', 'sugar', 'local', 'bit', 'review', 'still', 'nice', 'treat', 'co
me', 'lot', 'drink', 'differ', 'cant', 'say', 'keep', 'never', 'mani',
'two', 'could', 'dog', 'sinc', 'thing', 'sweet', 'know', 'add', 'who',
'brand', 'most', 'enjoy', 'need', 'favorit', 'packag', 'our', 'delici',
'way', 'think', 'work', 'again', 'made', 'their', 'purchas', 'then', 'o
ver', 'high', 'give', 'mix', 'want', 'look', 'bought', 'first', 'box',
'found', 'day', 'bag', 'after', 'recommend', 'which', 'better', 'now',
'year', 'ani', 'ive', 'well', 'were', 'what', 'there', 'food', 'becau
s', 'price', 'littl', 'even', 'store', 'been', 'tea', 'also', 'too', 'a
mazon', 'order', 'much', 'best', 'realli', 'dont', 'your', 'eat', 'abou
t', 'find', 'onli', 'some', 'would', 'time', 'buy', 'out', 'than', 'oth
er', 'more', 'had', 'will', 'has', 'get', 'when', 'make', 'from', 'it',
'can', 'all', 'tri', 'flavor', 'just', 'product', 'one', 'veri', 'the
m', 'use', 'great', 'good', 'these', 'love', 'tast', 'they', 'like', 'w
as', 'you', 'are', 'not', 'but', 'with', 'that', 'have', 'for', 'this',
'and', 'the']
```

co-occurrence matrix from top 3000 features with neighbour = 5

In [11]:
```
n_neighbor = 5
occ_matrix = np.zeros((3000,3000))
```

```
for row in tqdm(preprocessed):
    words_in_row = row.split()
    for index,word in enumerate(words_in_row):
        if(word in top_features):
            for j in range(max(index-n_neighbor,0),min(index+n_neighbor
,len(words_in_row)-1) + 1):
                if(words_in_row[j] in top_features):
                    occ_matrix[top_features.index(word),top_features.in
dex(words_in_row[j])] += 1
                else:
                    pass
        else:
            pass
```

```
100%|██████████| 100000/100000 [1:17:26<00:00, 21.52it/s]
```

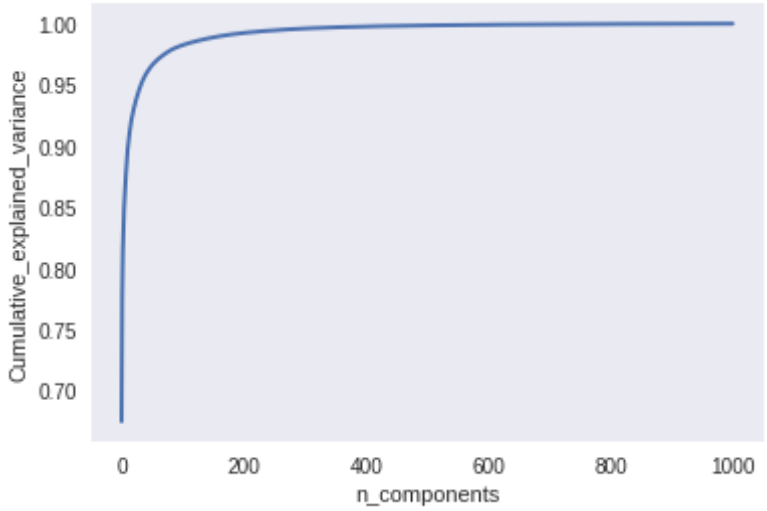
In [0]:
```
np.count_nonzero(occ_matrix)
#occ_matrix.size
```

Out[0]:
```
2128096
```

TrunctedSVD with n_components 1000 for finding maximum varience to compute n_components

In [0]:
```
from sklearn.decomposition import TruncatedSVD
from sklearn.preprocessing import StandardScaler
tsvd = TruncatedSVD(n_components = 1000)
svd = tsvd.fit_transform(occ_matrix)

percentage_var_explained = tsvd.explained_variance_ / np.sum(tsvd.expla
ined_variance_);
cum_var_explained = np.cumsum(percentage_var_explained)
plt.figure(figsize=(6, 4))

plt.clf()
plt.plot(cum_var_explained, linewidth=2)
plt.axis('tight')
```

```
plt.grid()

plt.xlabel('n_components')
plt.ylabel('Cumulative_explained_variance')
plt.show()
```



we got optimal n_componets 100 from the maximum cummlative varience from Above graph

so from n_components 100 find SVD

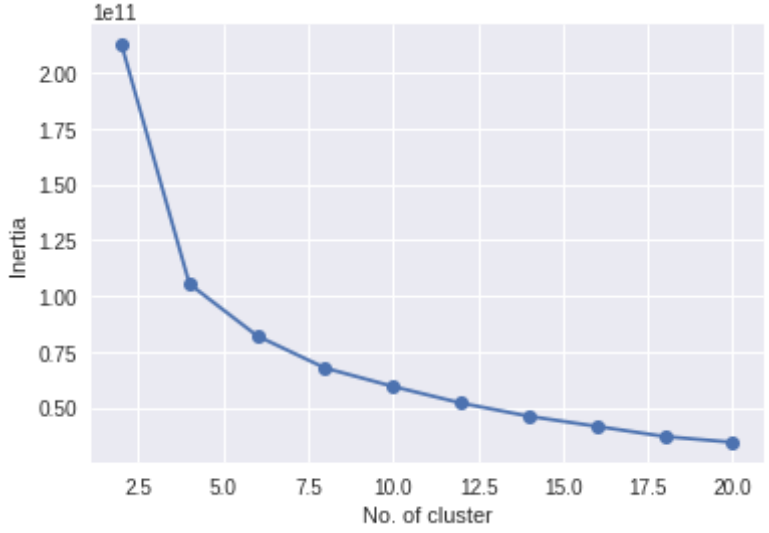
In [0]:
```
tsvd = TruncatedSVD(n_components = 100)
svd = tsvd.fit_transform(occ_matrix)
```

In [0]:
```
svd.shape
```

Out[0]:
```
(3000, 100)
```

Kmeans Clustering on SVD Data

In [0]:
```
clusters = [2,4,6,8,10,12,14,16,18,20]
from sklearn.cluster import KMeans
dic = {}
for i in clusters:
    clus = KMeans(n_clusters = i)
    clus.fit(svd)
    dic[i] = clus.inertia_

plt.plot(list(dic.keys()), list(dic.values()),'-o')
plt.xlabel("No. of cluster")
plt.ylabel("Inertia")
plt.show()
```



taken Optimal Cluster as 7

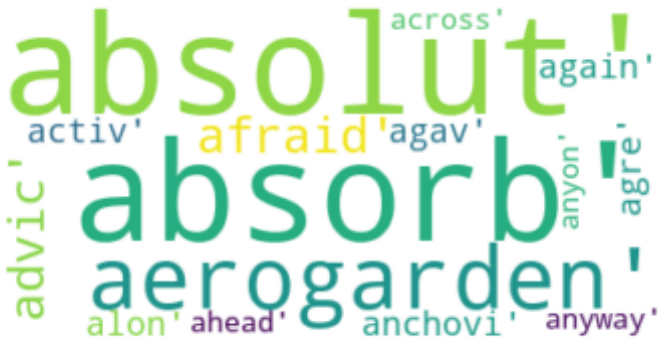
In [0]:
```
optimal_k = KMeans(n_clusters = 10)
p = optimal_k.fit_predict(svd)
```

In [0]:
```
from collections import Counter
a=optimal_k.labels_
type(a)
Counter(a.tolist())
```

Printing WordClouds for each Cluster

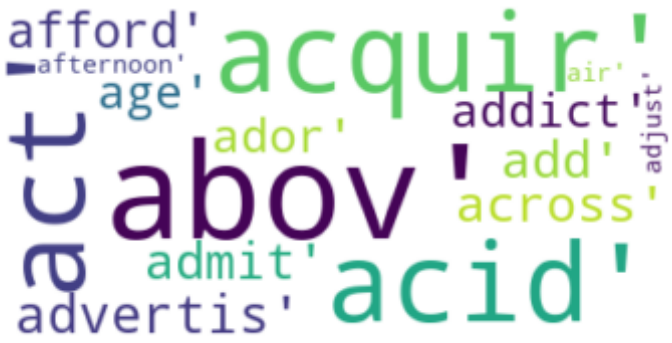
In [0]:
```
from wordcloud import WordCloud, STOPWORDS

stopwords_t = set(STOPWORDS)

centroids = optimal_k.cluster_centers_.argsort() # function for printin
g top 30 feature names with each cluster

terms = tf_idf_vect.get_feature_names()

list1 = []

for i in range(10):
    print("Cluster %d:" % i, end='')
    for j in centroids[i, :15]:
        list1.append(terms[j])
    wc = WordCloud(background_color="white", max_words=len(str(list1)),
 stopwords=stopwords_t)
    wc.generate(str(list1))
    print("Word Cloud for KMeans Cluster:", i)
    plt.imshow(wc, interpolation='bilinear')
    plt.axis("off")
    plt.show()
    list1.clear()
```

```
Cluster 0:Word Cloud for KMeans Cluster: 0
```



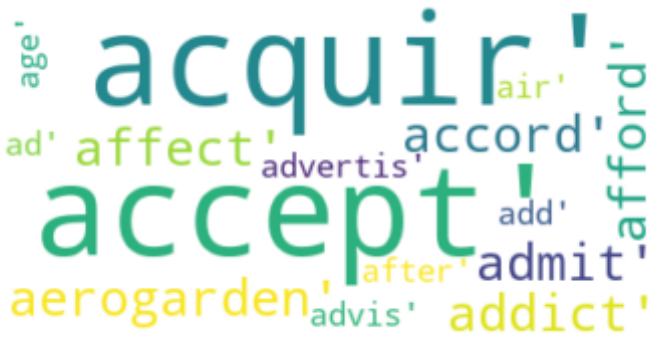
Cluster 1:Word Cloud for KMeans Cluster: 1

Cluster 2:Word Cloud for KMeans Cluster: 2

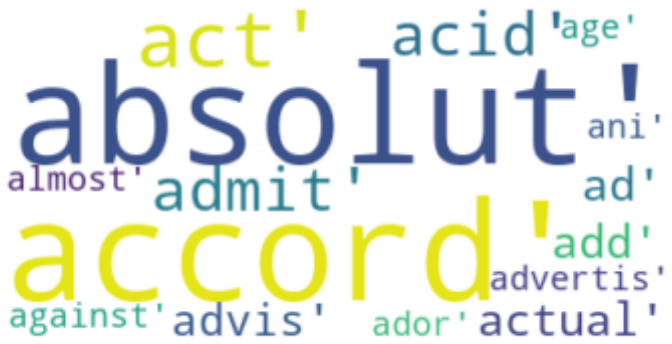


Cluster 3:Word Cloud for KMeans Cluster: 3

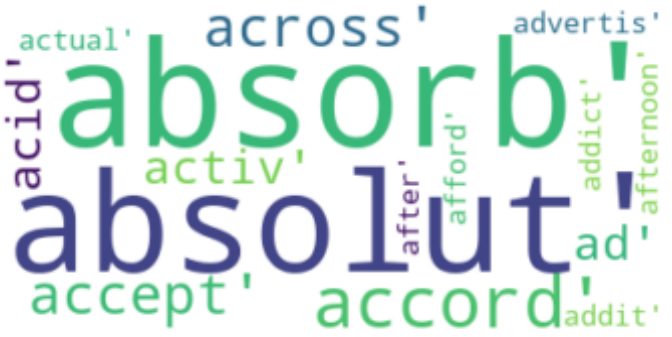
Cluster 4:Word Cloud for KMeans Cluster: 4



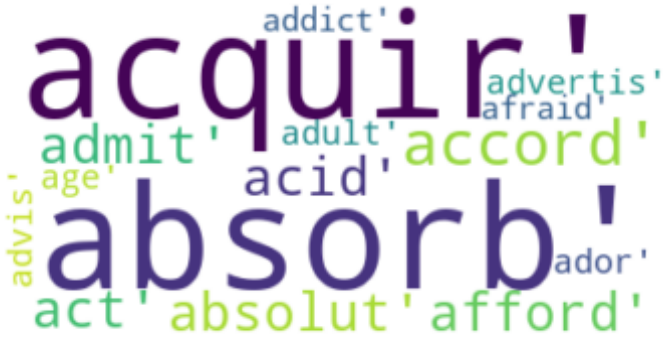
Cluster 5:Word Cloud for KMeans Cluster: 5



Cluster 6:Word Cloud for KMeans Cluster: 6

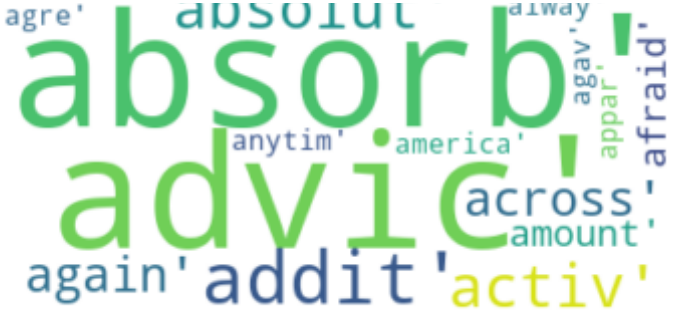


Cluster 7:Word Cloud for KMeans Cluster: 7



Cluster 8:Word Cloud for KMeans Cluster: 8

Cluster 9:Word Cloud for KMeans Cluster: 9



In [0]:

function that takes a word and returns the most similar words using cosine similarity between the vectors

In [0]:
```
from sklearn.metrics.pairwise import cosine_similarity
def similar_word_10(word):
    similarity = cosine_similarity(occ_matrix)
    word_vect = similarity[top_features.index(word)]
    print("Similar Word to",word)
    index = word_vect.argsort()[::-1][1:11]
    for j in range(len(index)):
        print((j+1),"Word",top_features[index[j]] ,"is similar to",word
,"\n")
```

In [0]:
```
similar_word_10(top_features[150])
```

```
Similar Word to nurs
1 Word and is similar to nurs

2 Word pump is similar to nurs

3 Word the is similar to nurs

4 Word eas is similar to nurs
```

```
5 Word for is similar to nurs

6 Word babi is similar to nurs

7 Word both is similar to nurs

8 Word mother is similar to nurs

9 Word relax is similar to nurs

10 Word aroma is similar to nurs
```

In [0]: