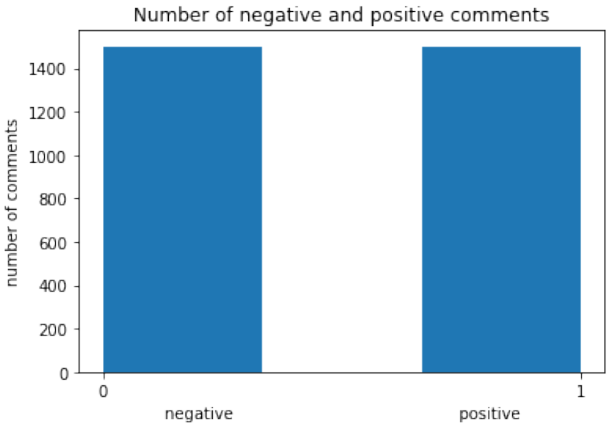
CS 5785: Homework 3 Write-up

Vianne Gao

10th November 2018

1. Sentiment Analysis for Online Reviews

(a) Data Overview



Labels are balanced according to the readme.txt file as well as the plot above. The first column of all three data files are parsed into a 1d numpy array (data) and the second column into another 1d numpy array (label).

(b) Preprocessing online comments

1. Lowercase all of the words: We want to focus on the vocabularies themselves for sentiment analysis. Lowercasing ensures that the case of the words do not affect the sentiment of the comment.
2. Stripping punctuation: Simplifies our data and allow us to focus on words in the comments.
3. Lemmatizing: Greatly reduces the dimensionality and sparcity of our data without losing much information.
4. Remove stop words: Another way to reduce dimensionality of our data and put more emphasis on words providing more sentiment information.

(d) Bag of Words model

Why we do not loop through the test set when building our dictionary:
If we loop through the test set when creating the word dictionary, and if there exists words present in the test set but not in the train set, then the feature will be 0 for all train samples. Hence, such word features are not contributing to classification.

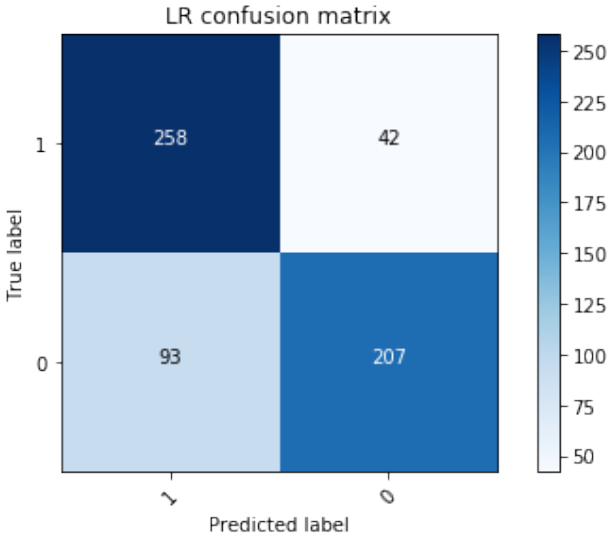
-Review 1: good case excel valu
$[0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, ..., 0, 0, 0]$
-Review 2: great jawbon
$[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, ..., 0, 0, 0]$

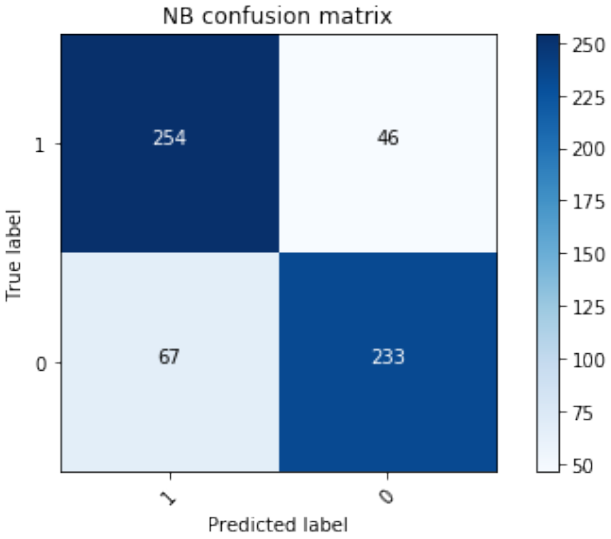
(e) Post-processing

L1 normalization was performed to transform the feature vector such that it represents the frequency of each word in the review. This allows a more direct comparison between reviews.

(f) Sentiment prediction

- **Logistic Regression Accuracy**: 0.775 - **Bernoulli NB accuracy:** 0.812
- **Confusion matrix:**





Most significant words: great, bad, love, excel, good, nice, best, poor, wast, delici

Most significant words: movi,bad,phone,time,would,great,good,film,phone,love

The Bernoulli Naive Bayes model has a higher accuracy than Logistic Regression model. Specifically, the false negative rate is lower in the Naive Bayes model.

(g) **N-gram model**

-Review 1: good case excel valu

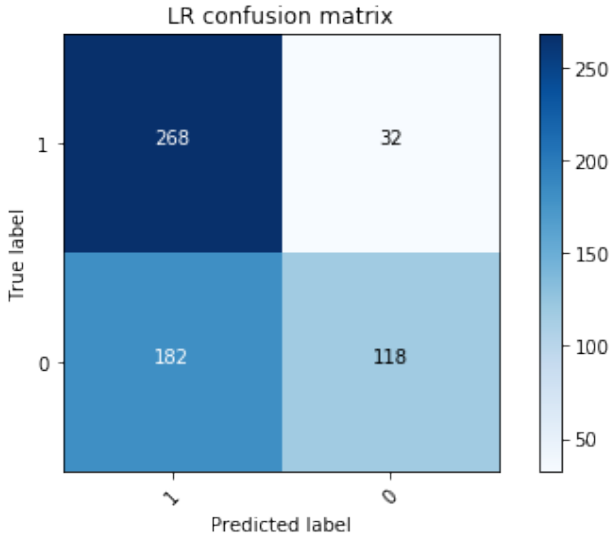
$[0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, ..., 0, 0]$

-Review 2: great jawbon

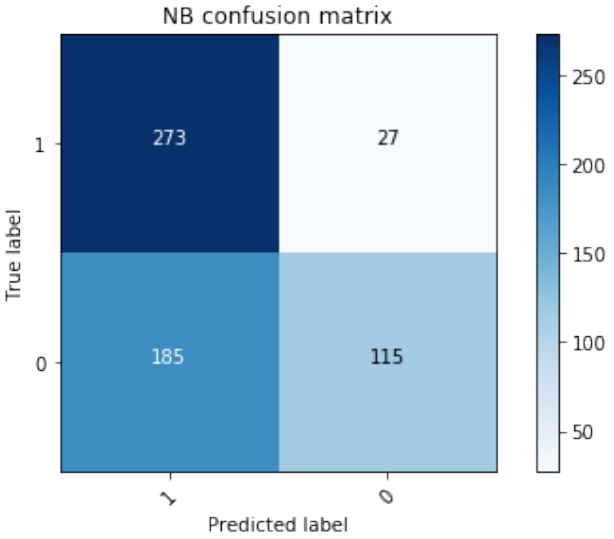
$[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ..., 0, 0, 0]$

- **Logistic Regression Accuracy**: 0.643 - **Bernoulli NB accuracy:** 0.647
- **Confusion matrix:**

The Bernoulli Naive Bayes model has a slightly higher accuracy than Logistic Regression model. Using the N-gram model results in a high false-positive rate in both classifiers.



Most significant words: work great, highli recommend,wast time,great phone,wast money,disappoint,great product,10 10,food good,easi use

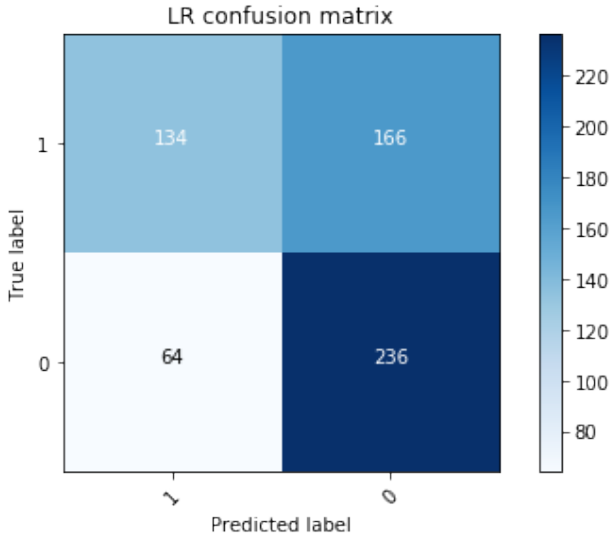


Most significant words: work great go back,wast time,custom servic,disappoint,wast money,work great, highli recommend, one best, sound qualiti, work well

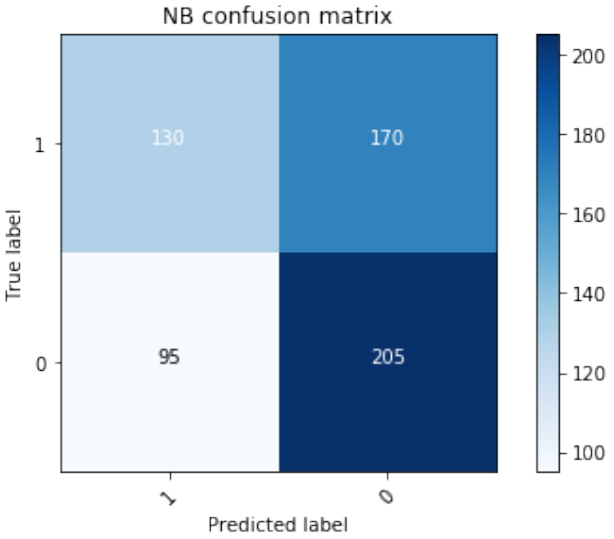
(h-f) **PCA with Bag of Words model**

———————— **PC = 10** ————————-:

- **Logistic Regression Accuracy**: 0.617
- **Bernoulli NB accuracy:** 0.558
- **Confusion matrix:**



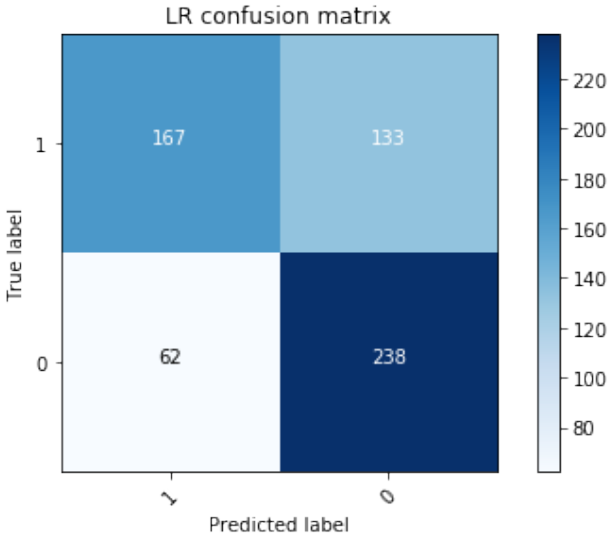
Most significant words: way, plug, say, go, volum, fun, case, one, convert, sever

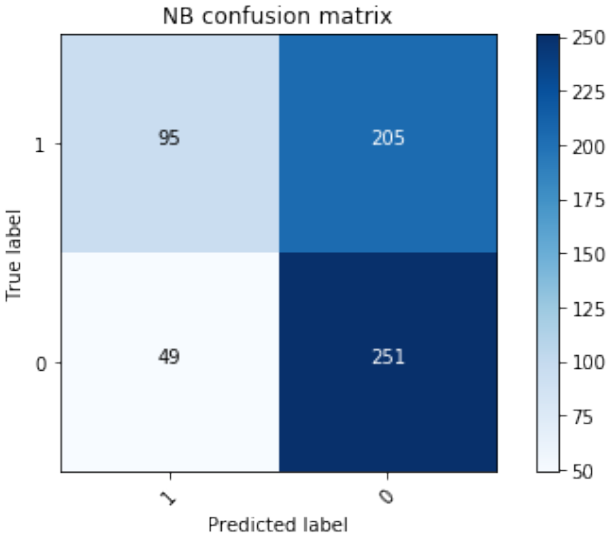


Most significant words: plug, convert,case,go,valu, good,excel, unless, way, us

———————— **PC = 50** ————————-:

- **Logistic Regression Accuracy**: 0.675
- **Bernoulli NB accuracy:** 0.577
- **Confusion matrix:**



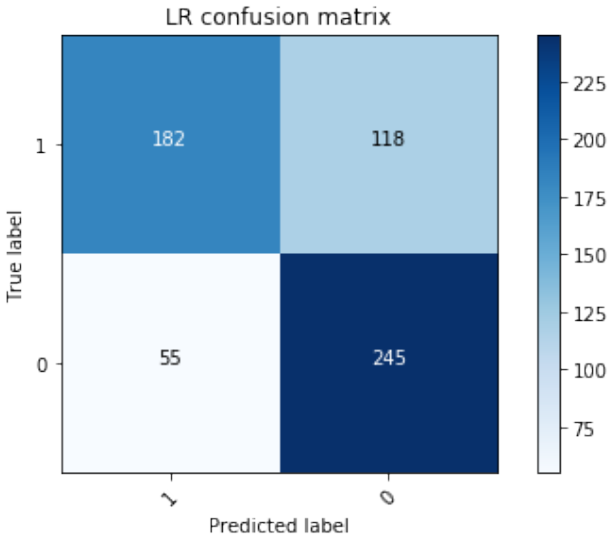


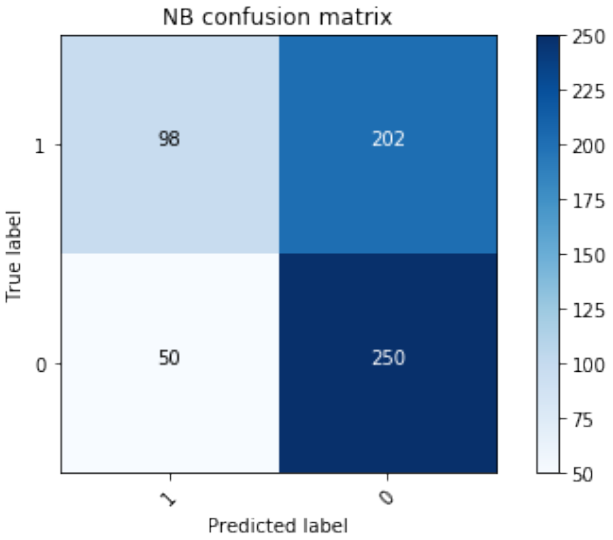
Most significant words: way, plug, go, case, convert, valu, us, excel, good,unless

Most significant words: plug, convert,mic, 45, case, valu, minut, must, hundr, say, imagin

——————— **PC = 100** ———————-:

- **Logistic Regression Accuracy**: 0.712
- **Bernoulli NB accuracy:** 0.580
- **Confusion matrix:**





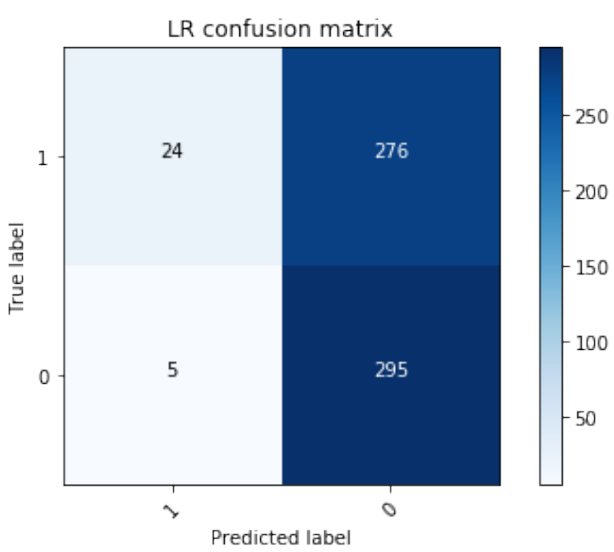
Most significant words: way, plug, say,go, volum, fun, case, one, convert, severe

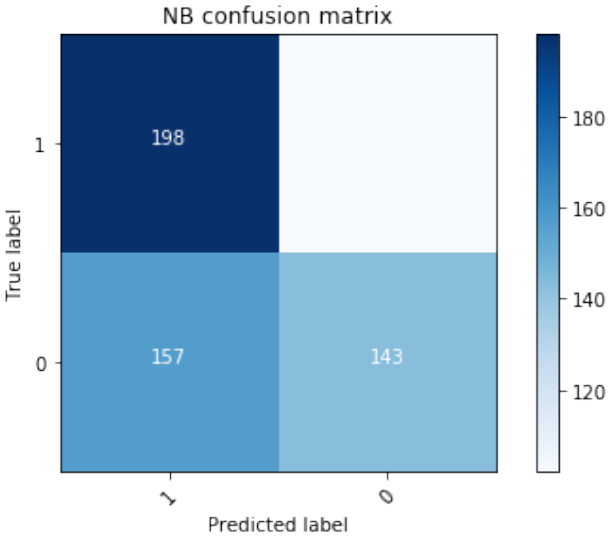
Most significant words: plug, convert,mic, 45, case, valu, minut, must, hundr, say, imagin

(h-g) **PCA with N-gram model**

———————— **PC = 10** ————————-:

- **Logistic Regression Accuracy**: 0.532
- **Bernoulli NB accuracy:** 0.600
- **Confusion matrix:**



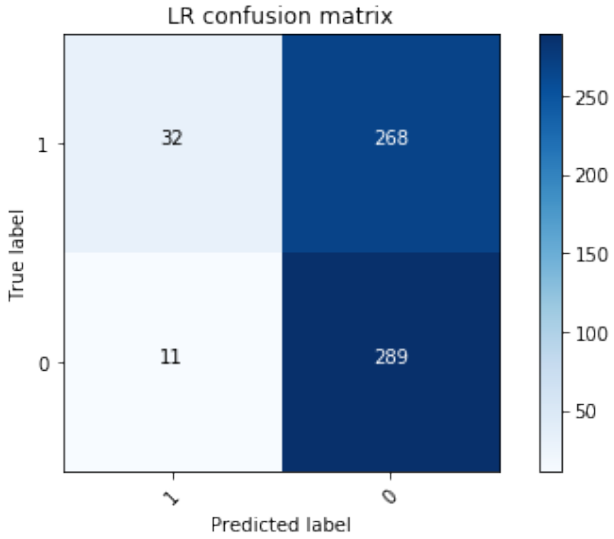


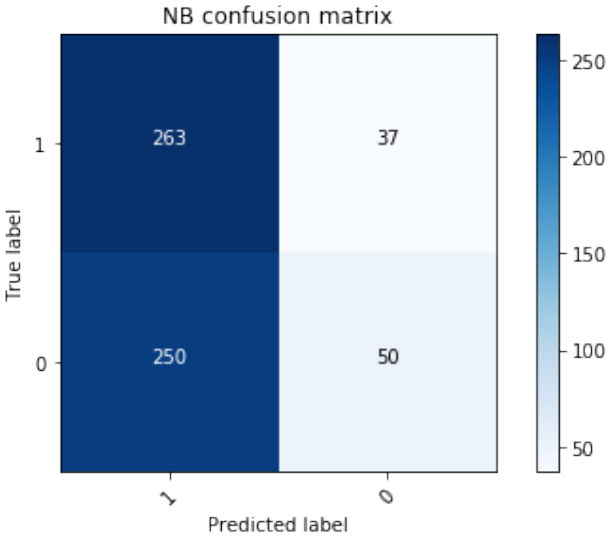
Most significant words: plug us, excel valu, good case, go convert, way plug, us unless, tie charger, unless go, great jawbon, case excel

Most significant words: plug us, unless go, case excel, great jawbon, excel valu,tie charger,good case, go convert, way plug, us unless

———————— **PC = 50** ————————-:

- **Logistic Regression Accuracy**: 0.535
- **Bernoulli NB accuracy:** 0.522
- **Confusion matrix:**



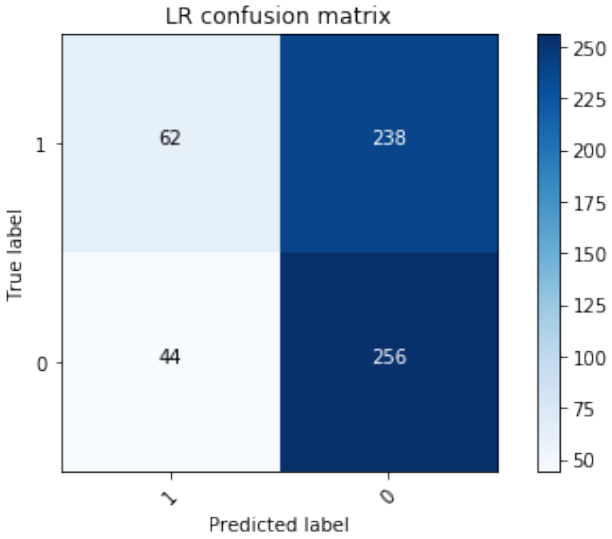


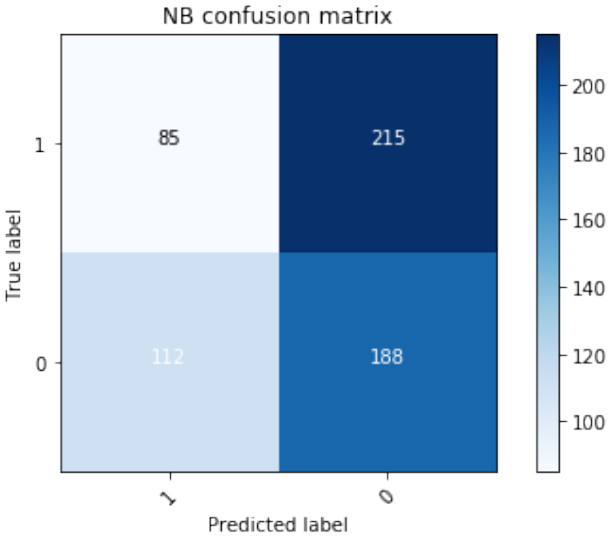
Most significant words: plug us, excel valu, good case, mere 5, go convert, way plug, us unless, send one, tie charger, convers last

Most significant words: plug us, unless go, case excel, convers last, 45 minut, line right, great jawbon, get decent, plug get, mic great

——————— **PC = 100** ———————-:

- **Logistic Regression Accuracy**: 0.530 - **Bernoulli NB accuracy:** 0.455
- **Confusion matrix:**





Most significant words: plug us, excel valu, good case, mere 5, go convert, way plug, us unless, send one, tie charger, convers last

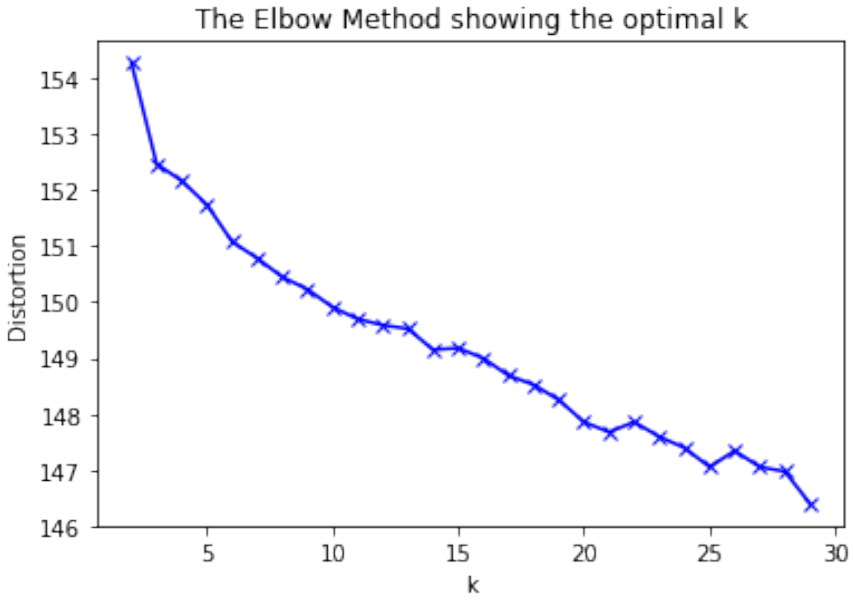
Most significant words: plug us, unless go, case excel, convers last, 45 minut, line right, great jawbon, get decent, plug get, mic great

(i) For naive bayes, bag of words model gave the best AUC (0.88) and accuracy (0.812).
For linear regression, PCA on bag of words (700 features) gave the best accuracy at 0.793, but bag of words gave the best AUC at 0.872.
Another thing to consider is the speed of our classifier. Performing PCA on the bag of words model can speed up classification. We obtained 0.86 AUC and 0.77 accuracy using 300 features using linear regression, which is quite close to that obtained using 3500 features with naive bayes.
Another thing to note in linear regression is that after conducting PCA, the true negative rate increases while the true positive rate decreases. Hence, the desirability of each model depends on our goal as well.

People often use keywords like great, excellent, good, best, love etc. to express positive, and poor, bad, waste etc. to express negative sentiment. However, there still are huge variations in vocabularies people use to express sentiment.

2. Clustering for Text Analysis

(a) Using the elbow method, we chose k=9.



Words associated with top components and topics closest to the cluster center:

Cluster 2:

['population'] ['significant'] ['significantly'] ['reports'] ['responses'] ['mean'] ['test'] ['email'] ['analysis'] ['populations']

['"Selectivity for 3D Shape That Reveals Distinct Areas within Macaque Inferior Temporal Cortex"']

['"Nonrandom Extinction and the Loss of Evolutionary History"']

['"Mirror-Image Confusion in Single Neurons of the Macaque Inferotemporal Cortex"']

['"An Empirical Assessment of Taxic Paleobiology"']

['"Abolition and Reversal of Strain Differences in Behavioral Responses to Drugs of Abuse after a Brief Experience"']

['"Promiscuity and the Primate Immune System"']

['"Language Discrimination by Human Newborns and by Cotton-Top Tamarin Monkeys"']

['"Natural Selection and Parallel Speciation in Sympatric Sticklebacks"']

['"Reversal of Antipsychotic-Induced Working Memory Deficits by Short-Term Dopamine D1 Receptor Stimulation"']

['"High Direct Estimate of the Mutation Rate in the Mitochondrial Genome of Caenorhabditis elegans"']

Cluster 8:

['energy'] ['reports'] ['electron'] ['fig'] ['solid'] ['optical'] ['shows'] ['dependence'] ['sample'] ['magnetic']

['"The Formation of Chondrules at High Gas Pressures in the Solar Nebula"']

['"A Monoclinic Post-Stishovite Polymorph of Silica in the Shergotty Meteorite"']

['"Synthesis and Characterization of Helical Multi-Shell Gold Nanowires"']

['"Ambipolar Pentacene Field-Effect Transistors and Inverters"']

['"A Stable Bicyclic Compound with Two Si=Si Double Bonds"']

['"Xenon as a Complex Ligand: The Tetra Xenono Gold(II) Cation in ¡latex¿$AuXe_4^{2+}(Sb_2F_{11}^-)_2$¡/latex¿"']

['"Direct Condensation of Carboxylic Acids with Alcohols Catalyzed by Hafnium(IV) Salts"']

['"Atomic Layer Deposition of Oxide Thin Films with Metal Alkoxides as Oxygen Sources"']

['"A Cyclic Carbanionic Valence Isomer of a Carbocation: Diphosphino Analogs of Diaminocarbocations"']

['"Discovery of a Basaltic Asteroid in the Outer Main Belt"']

Cluster 5:

['mail'] ['compass'] ['author'] ['page'] ['issue'] ['news'] ['sciences'] ['policy'] ['department'] ['edu']
['"Algorithmic Gladiators Vie for Digital Glory"']

['"Reopening the Darkest Chapter in German Science"']

['"National Academy of Sciences Elects New Members"']

['"Corrections and Clarifications: Unearthing Monuments of the Yarmukians"']

['"Corrections and Clarifications: Charonś First Detailed Spectra Hold Many Surprises"']

['"Corrections and Clarifications: A Short Fe-Fe Distance in Peroxodiferric Ferritin: Control of Fe Substrate versus Cofactor Decay?"']

['"Heretical Idea Faces Its Sternest Test"']

['"Archaeology in the Holy Land"']

['"Corrections and Clarifications: One Hundred Years of Quantum Physics"']

['"Corrections and Clarifications: Biotech Research Proves a Draw in Canada"']

Cluster 3:

['climate'] ['ocean'] ['atmospheric'] ['atmosphere'] ['sea'] ['global'] ['records'] ['atlantic'] ['record'] ['variability']

['"Population Dynamical Consequences of Climate Change for a Small Temperate Songbird"']

['"Reconstruction of the Amazon Basin Effective Moisture Availability over the past 14,000 Years"']

['"Frozen Methane Escapes from the Sea Floor"']

['"Greenland Ice Sheet: High-Elevation Balance and Peripheral Thinning"']

['"The Causes of 20th Century Warming"']

['"Isotopic Evidence for Variations in the Marine Calcium Cycle over the Cenozoic"']

['"Glacial Climate Instability"']

['"Lessons for a New Millennium"']

['"Variable Carbon Sinks"']

['"The Role of the Southern Ocean in Uptake and Storage of Anthropogenic Carbon Dioxide"']

Cluster 1:

['cells'] ['expression'] ['protein'] ['cell'] ['gene'] ['wild'] ['proteins'] ['mutant'] ['expressed'] ['control']

['"Requirement of NAD and SIR2 for Life-Span Extension by Calorie Restriction in Saccharomyces Cerevisiae"']

['"Suppression of Mutations in Mitochondrial DNA by tRNAs Imported from the Cytoplasm"']

['"Distinct Classes of Yeast Promoters Revealed by Differential TAF Recruitment"']

['"Efficient Initiation of HCV RNA Replication in Cell Culture"']

['"Negative Regulation of the SHATTERPROOF Genes by FRUITFULL during Arabidopsis Fruit Development"']

['"T Cell-Independent Rescue of B Lymphocytes from Peripheral Immune Tolerance"']

['"Patterning of the Zebrafish Retina by a Wave of Sonic Hedgehog Activity"']

['"Reduced Food Intake and Body Weight in Mice Treated with Fatty Acid Synthase Inhibitors"']

['"Coupling of Stress in the ER to Activation of JNK Protein Kinases by Transmembrane Protein Kinase IRE1"']

['"An Anti-Apoptotic Role for the p53 Family Member, p73, during Developmental Neuron Death"']

Cluster 0:

['earth'] ['geophys'] ['planet'] ['thermal'] ['depth'] ['material'] ['mantle'] ['composition'] ['crust'] ['surface']

['"Remobilization in the Cratonic Lithosphere Recorded in Polycrystalline Diamond"']

['"Extinct ¡latex¿^{129}I¡/latex¿ in Halite from a Primitive Meteorite: Evidence for Evaporite Formation in the Early Solar System"']

['"Evidence for Crystalline Water and Ammonia Ices on Plutoś Satellite Charon"']

['"African Hot Spot Volcanism: Small-Scale Convection in the Upper Mantle beneath Cratons"']

['"Solar Wind Record on the Moon: Deciphering Presolar from Planetary Nitrogen"']

['"Geodynamic Evidence for a Chemically Depleted Continental Tectosphere"']

['"Rutile-Bearing Refractory Eclogites: Missing Link between Continents and Depleted Mantle"']

['"Folds on Europa: Implications for Crustal Cycling and Accommodation of Extension"']

['"High Magma Storage Rates before the 1983 Eruption of Kilauea, Hawaii"']

['"Suppression of Rain and Snow by Urban and Industrial Air Pollution"']

Cluster 6:

['says'] ['researchers'] ['scientists'] ['people'] ['get'] ['year'] ['just'] ['say'] ['team'] ['national']

['"Information Technology Takes a Different Tack"']

['"Science Survives in Breakthrough States"']

['"Vaccine Studies Stymied by Shortage of Animals"']

['"For Fátheróf Abortion Drug, Vindication at Last"']

['"On a Slippery Slope to Mediocrity?"']

['"In Europe, Hooligans Are Prime Subjects for Research"']

['"Japanś Whaling Program Carries Heavy Baggage"']

['"Is AIDS in Africa a Distinct Disease?"']

['"New Science Chief Must Juggle Missions and Politics"']

['"Building a Disease-Fighting Mosquito"']

Cluster 7:

['protein'] ['binding'] ['proteins'] ['residues'] ['conserved'] ['structural'] ['amino'] ['acid'] ['domain'] ['cell']

['"Structure of Yeast Poly(A) Polymerase Alone and in Complex with 3′-dATP"']

['"Structure of Murine CTLA-4 and Its Role in Modulating T Cell Responsiveness"']

['"Selfish DNA in Protein-Coding Genes of Rickettsia"']

['"Candidate Taste Receptors in Drosophila"']

['"Structure of the S15,S6,S18-rRNA Complex: Assembly of the 30S Ribosome Central Domain"']

['"The Productive Conformation of Arachidonic Acid Bound to Prostaglandin Synthase"']

['"Atomic Structure of PDE4: Insights into Phosphodiesterase Mechanism and Specificity"']

['"Twists in Catalysis: Alternating Conformations of Escherichia coli Thioredoxin Reductase"']

['"Redox Signaling in Chloroplasts: Cleavage of Disulfides by an Iron-Sulfur Cluster"']

['"Structure of the Protease Domain of Memapsin 2 (b-Secretase) Complexed with Inhibitor"']

Cluster 4:

['energy'] ['theoretical'] ['matter'] ['motion'] ['momentum'] ['theory'] ['field'] ['velocity'] ['direction'] ['properties']

['"Gone with the Wind: The Origin of S0 Galaxies in Clusters"']

['"Generating Solitons by Phase Engineering of a Bose-Einstein Condensate"']

['"The Dark Halo of the Milky Way"']

['"Negative Poissonś Ratios for Extreme States of Matter"']

['"Quantum Criticality: Competing Ground States in Low Dimensions"']

['"Orbital Physics in Transition-Metal Oxides"']

['"The Galactic Center: An Interacting System of Unusual Sources"']

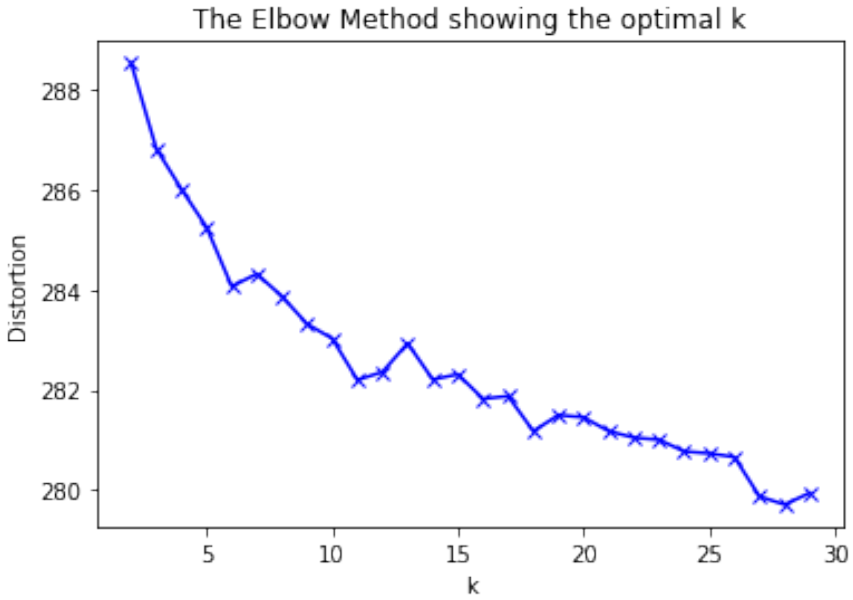
['"The Baryon Halo of the Milky Way: A Fossil Record of Its Formation"']

['"Whither the Future of Controlling Quantum Phenomena?"']

['"One Hundred Years of Quantum Physics"']

The algorithm captures the field of study and style of Science papers. Such algorithm is useful for categorizing and organizing papers in a database and improve the result of search engines.

(b) Using the elbow method, we chose k=5.



Titles associated with top components and words closest to the cluster center:

Cluster 4:

['aptamers'] ['dnag'] ['trxr'] ['rory'] ['lcts'] ['proteorhodopsin'] ['ag7'] ['nompc'] ['lg268'] ['neas']

['"Into the Forbidden Zone"']

['"Clues from a Shocked Meteorite"']

['"Influences of Dietary Uptake and Reactive Sulfides on Metal Bioavailability from Aquatic Sediments"']

['"Nonavian Feathers in a Late Triassic Archosaur"']

['"Ambipolar Pentacene Field-Effect Transistors and Inverters"']

['"The Formation of Chondrules at High Gas Pressures in the Solar Nebula"']

['"National Academy of Sciences Elects New Members"']

['"A Monoclinic Post-Stishovite Polymorph of Silica in the Shergotty Meteorite"']

['"The Chi-Chi Earthquake Sequence: Active, Out-of-Sequence Thrust Faulting in Taiwan"']

['"Synthesis and Characterization of Helical Multi-Shell Gold Nanowires"']

Cluster 3:

['expectancy'] ['celera'] ['intelligence'] ['managers'] ['income'] ['teachers'] ['doe'] ['mosquitoes'] ['mosquito'] ['essay']

['"Presidential Forum: Gore and Bush Offer Their Views on Science"']

['"Something to Be Done: Treating HIV/AIDS"']

['"A Mouse Chronology"']

['"Ecologists on a Mission to Save the World"']

['"Silent No Longer: ́Model Minority ́Mobilizes"']

['"Infectious History"']

['"Eligibility for CSEM Scholarships"']

['"Balancing the Collaboration Equation"']

['"Ground Zero: AIDS Research in Africa"']

['"Biological Control of Invading Species"']

Cluster 2:

['www'] ['approximation'] ['angular'] ['finite'] ['coherent'] ['nonlinear'] ['periodic'] ['regime'] ['calculation'] ['diffraction']

['"NEAR at Eros: Imaging and Spectral Results"']

['"The Atom-Cavity Microscope: Single Atoms Bound in Orbit by Single Photons"']

['"Advances in the Physics of High-Temperature Superconductivity"']

['"Subduction and Slab Detachment in the Mediterranean-Carpathian Region"']

['"Internal Structure and Early Thermal Evolution of Mars from Mars Global Surveyor Topography and Gravity"']

['"The Formation and Early Evolution of the Milky Way Galaxy"']

['"Quantum Criticality: Competing Ground States in Low Dimensions"']

['"Earthś Core and the Geodynamo"']

['"Sediments at the Top of Earthś Core"']

['"Experiments and Simulations of Ion-Enhanced Interfacial Chemistry on Aqueous NaCl Aerosols"']

Cluster 1:

['whats'] ['thing'] ['researcher'] ['didnt'] ['doesnt'] ['hopes'] ['got'] ['plans'] ['biologist'] ['getting']

['"Atom-Scale Research Gets Real"']

['"Meltdown on Long Island"']

['"Help Needed to Rebuild Science in Yugoslavia"']

['"A Mouse Chronology"']

['"Clones: A Hard Act to Follow"']

['"Designer Labs: Architecture Discovers Science"']

['"Id́ like to See America Used as a Global Lab"']

['"Creationś Seventh Day"']

['"Soft Moneyś Hard Realities"']

['"Ecologists on a Mission to Save the World"']

Cluster 0:

['immunoblotting'] ['immunoblot'] ['immunoprecipitated'] ['plasmids'] ['polyacrylamide'] ['lysates'] ['immunoglobulin'] ['bovine'] ['wildtype'] ['phosphorylated']

['"Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator of p53-Induced Apoptosis"']

['"Central Role for G Protein-Coupled Phosphoinositide 3-Kinase g in Inflammation"']

['"Kinesin Superfamily Motor Protein KIF17 and mLin-10 in NMDA Receptor-Containing Vesicle Transport"']

['"Role of the Mouse ank Gene in Control of Tissue Calcification and Arthritis"']

['"Regulated Cleavage of a Contact-Mediated Axon Repellent"']

['"Positional Syntenic Cloning and Functional Characterization of the Mammalian Circadian Mutation tau"']

['"Dual Signaling Regulated by Calcyon, a D1 Dopamine Receptor Interacting Protein"']

['"Requirement of JNK for Stress-Induced Activation of the Cytochrome c-Mediated Death Pathway"']

['"Regulation of STAT3 by Direct Binding to the Rac1 GTPase"']

['"Function of PI3Kg in Thymocyte Development, T Cell Activation, and Neutrophil Migration"']

This algorithm can help to connect different documents discussing the same topics. This can be useful when building search engines.

Clustering documents outputs those using similar keywords, hence most clusters have similar styles. Clustering keywords outputs words that often appear together in documents. The documents may be of different genres, but they most likely discuss the same topics.

3. EM Algorithm and Implementation

(a) Kmeans and EM

① We can take σ^2 to be 0, hence π_k becomes 1 for the most probable class and 0 for the rest. Suppose we have n samples.

② E-step (calculate the responsibility of cluster k on sample i)

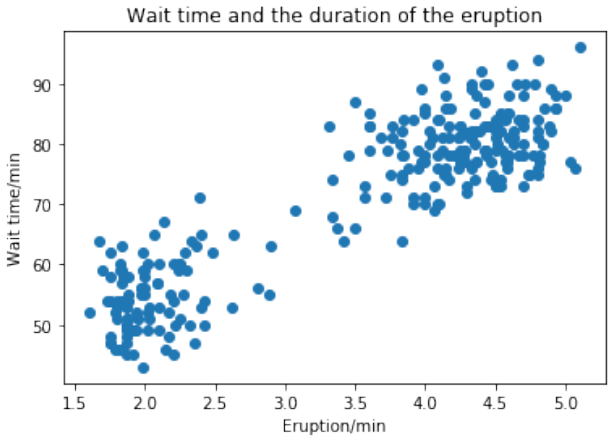
$$\gamma_{ik} = \begin{cases} 1, & \text{if } k = \arg\min_j (x_i - \bar{x}_j)^2 \\ 0, & \text{otherwise} \end{cases}$$

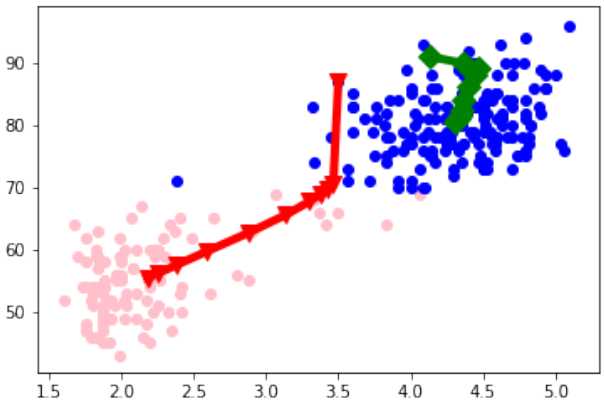
③ M-step (calculate the mean of each cluster)

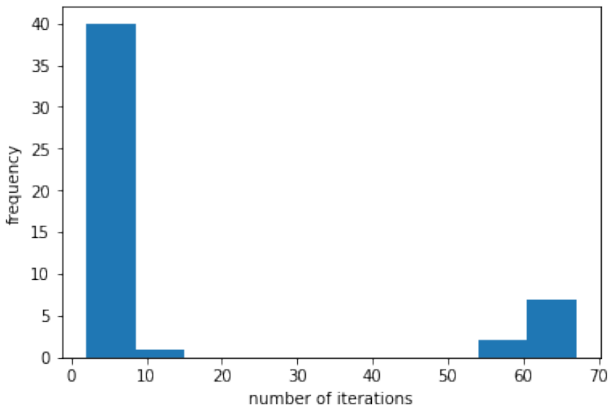
$$\begin{cases} \mu_k = \dfrac{\sum_{i=1}^{n} \gamma_{ik} x_i}{\sum_{i=1}^{n} \gamma_{ik}} \\ \sigma_k^2 = 0 \quad \text{by assumption.} \end{cases}$$

Repeat until there is no change in assignment.

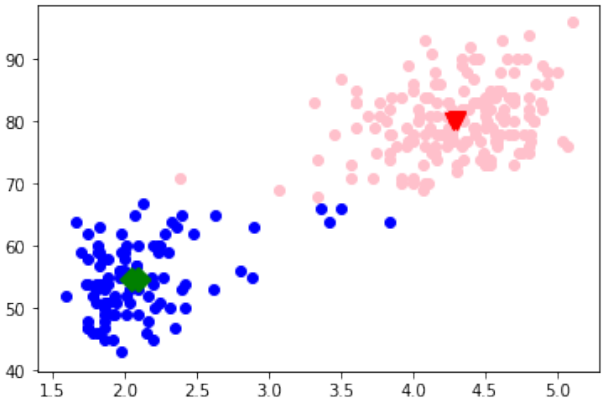
(b) Visualizing the data

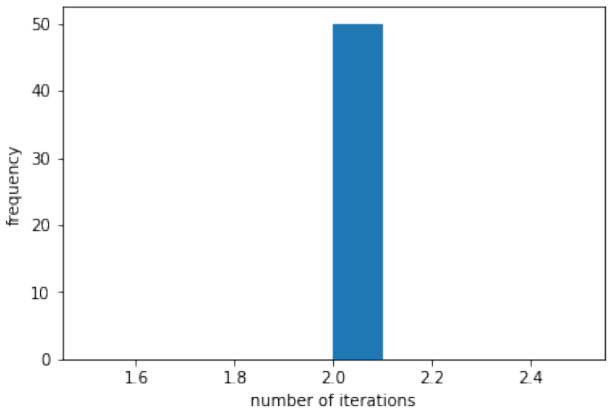






Random initialization





Kmeans initialization

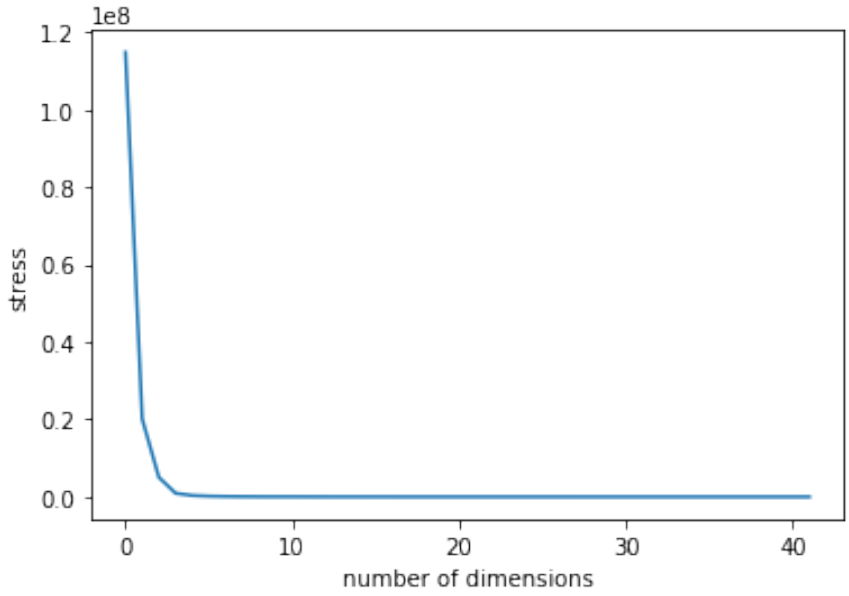
Although the clustering do not differ much between the two methods of initialization, the number of iterations needed until convergence is significantly lower when we initialize using kmeans.

4. Multidimensional scaling for genetic population differences

(a-i) The algorithm assumes that Nei's distance can be represented by euclidean distance in a 2-dimensional space. Also, this assumes that Nei's distance has no other information such as orientation. For example, MDS can fail if four or more data points all have the same pair-wise distance. It is impossible to reflect this relationship in a 2d space using euclidean distance. In this case, we may need to sacrifice global configuration to satisfy local configurations during the scaling process.

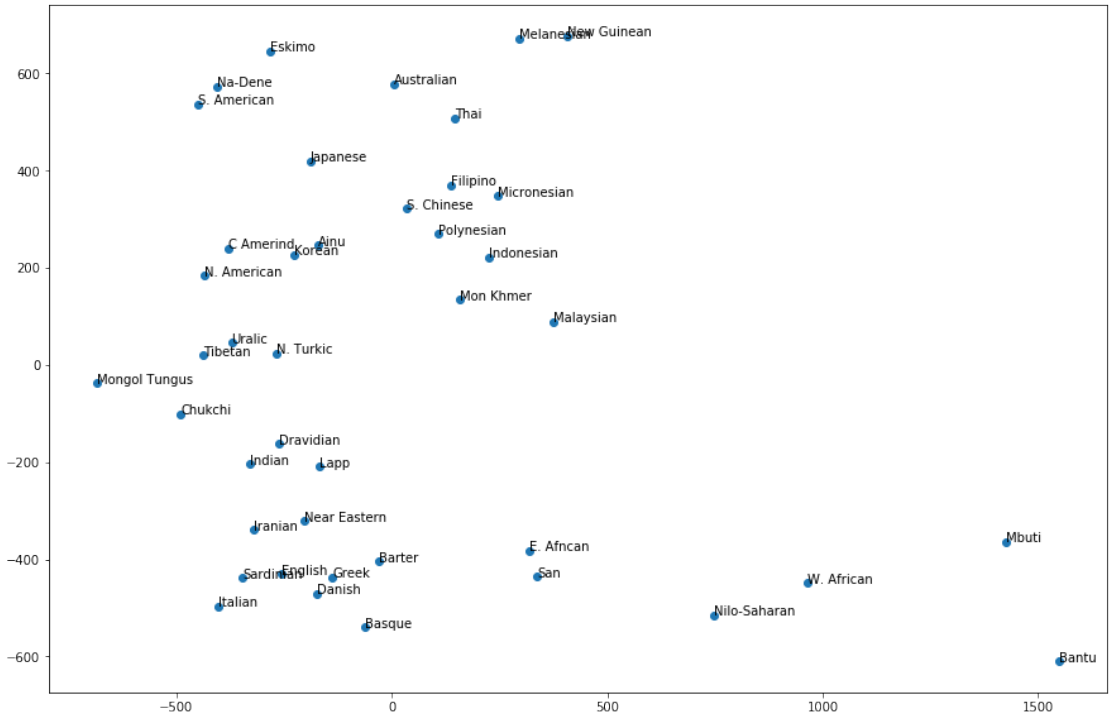
We can measure the amount of information lost by calculating the squared sum of differences between the Nei's distance and euclidean distance between every pair of points.

(a-ii) We measured the stress of our scaling as m increase.



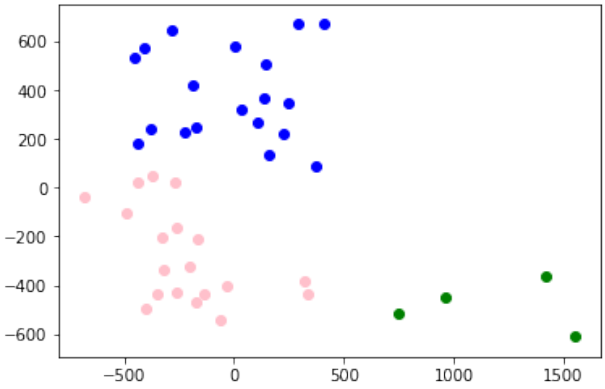
From the figure above, m=3 seems like a good fit.

(a-iii) 2D embedding and Kmeans



2D embedding of datapoints

(b) 2D embedding and Kmeans



Kmeans clustering result.

Three clusters:

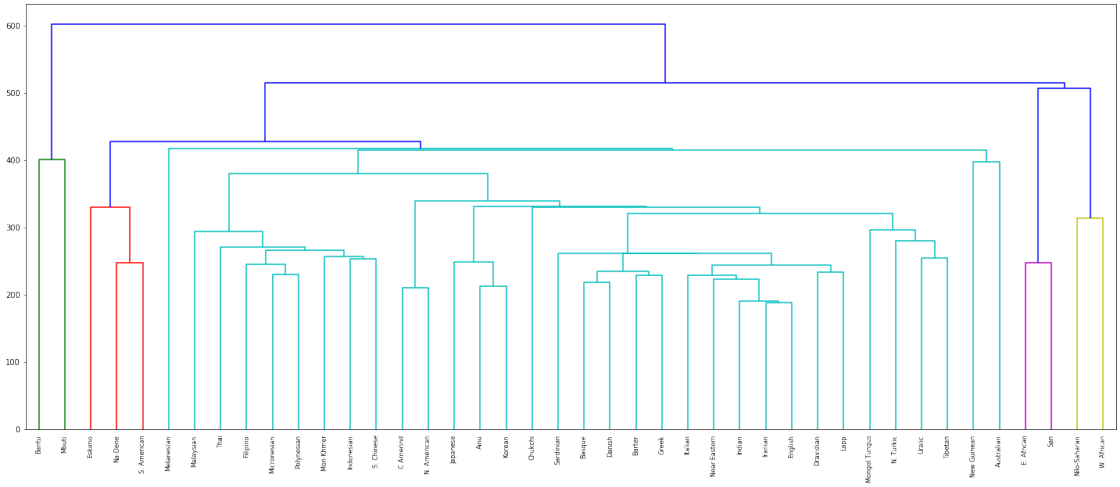
['E. Afncan' 'San' 'Barter' 'Indian' 'Iranian' 'Near Eastern' 'Uralic' 'Dravidian' 'Mongol Tungus' 'Tibetan' 'N. Turkic' 'Basque' 'Lapp' 'Sardinian' 'Danish' 'English' 'Greek' 'Italian' 'Chukchi']

['Ainu' 'Japanese' 'Korean' 'Mon Khmer' 'Thai' 'Indonesian' 'Malaysian' 'Filipino' 'S. Chinese' 'C Amerind' 'Eskimo' 'Na-Dene' 'N. American' 'S. American' 'Melanesian' 'Micronesian' 'Polynesian' 'New Guinean' 'Australian']

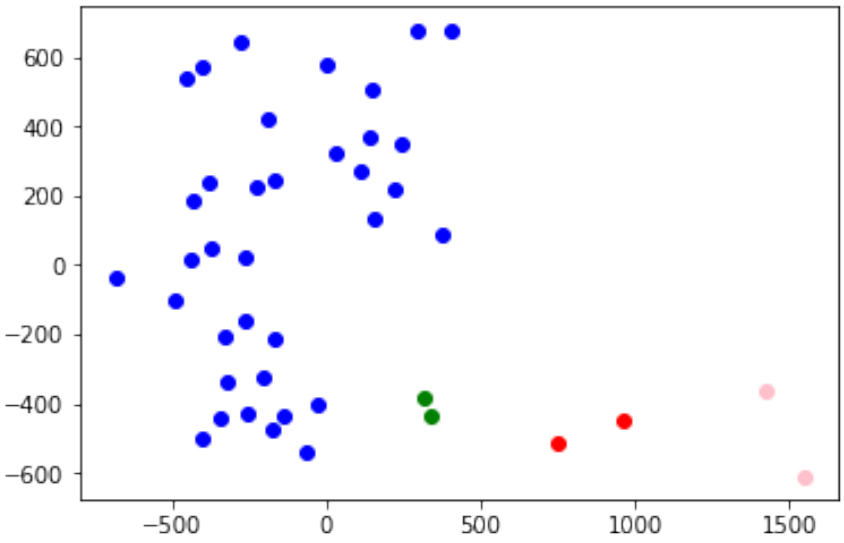
['Bantu' 'Nilo-Saharan' 'W. African' 'Mbuti']

Do not completely agree with the clustering. The clusters reflect populations on the same continent quite well, but not always. For instance, You would expect Tibetan and Mongol Tungus to be in the same cluster as the other Asian countries. Also, you would expect Australians to be clustered more closely with the other European populations. Such embedding does not preserve the global structure of the dissimilarities.

(c) Comparing hierarchical clustering with K-Means



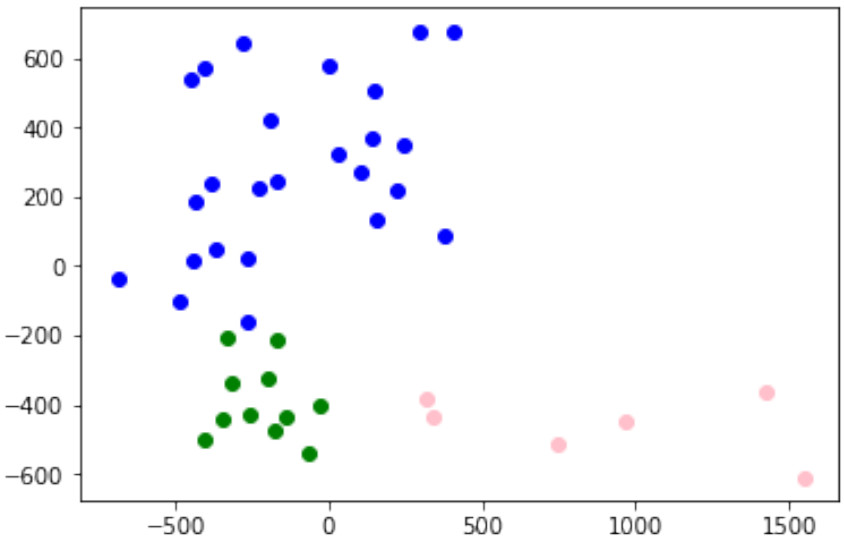
dendrogram of datapoints



Result of slicing the dendrogram.

Compared to kmeans, hierarchical clustering results in a huge cluster that are relatively more similar to each other, and three tiny clusters that are more different from the rest. Note that we are performing hierarchical clustering on the original distance matrix, while kmeans was performed on the embedded datapoints.

(d) K-medoids and kmeans



Kmedoids clustering result.

There isn't a significant difference in the clustering result. One thing to note is that Kmedoids clustered the more different points (those that were not in the big cluster generated by hierarchical clustering) into one cluster, while kmeans only clustered the four furthest points into one cluster.