

CS5785 Homework 3

Fei Li (fl392)

Shunzhe Yu (sy679)

November 10, 2018

Problem 1. Sentiment Analysis of Online Reviews.

(a) The labels are balanced (500 : 500 for each file). I read each file and process it line by line and put the data into a Pandas DataFrame (1000 rows). Then I concat the 3 DataFrames into one DataFrame (3000 rows) with hierarchical indexing (amazon, imdb and yelp).

(b) I picked all preprocessing strategies listed in the instruction and some other strategies I found at <https://machinelearningmastery.com/clean-text-machine-learning-python/>. The reasons are as follows.

- expand contractions
This strategy simply performs replacement to contractions, such as I'm to I am, what's to what is and etc. This is very useful for dimensionality reduction.
- lowercase all of the words
This strategy also reduces dimensionality, since we should treat the same word with different cases as the same entry in the word vector.
- strip punctuation
Punctuation cannot clearly reflect the sentiment of a reviews. eg. "!" can be used in both positive and negative sentiment.
- strip numbers
Numbers are not useful in sentiment analysis either.
- strip the stop words
The stop words contain no meaning in English, so we can omit them to reduce the dimensionality of word vectors.
- stemming and lemmatization
Stemming and lemmatization reduce each word to its base or root, which also reduce the dimensionality of word vectors. I use SnowballStemmer and WordNetLemmatizer in this problem.

(d) Each feature vectors has 3504 length, i.e. representing 3504 words. The first two reviews' feature vectors are as follows.

```
print training_set.iloc[0]
print training_set.iloc[1]

sentence                                way plug u unless go convert
score                                                                0
feature_vector    [1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
Name: 0, dtype: object
sentence                                tie charger convers last minutesmajor problem
score                                                                0
feature_vector    [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, ...
Name: 3, dtype: object
```

(e) I use log-normalization. First, I tried all 4 strategies and it performs best. Second, I find the reason to use log at <http://onlinestatbook.com/2/transformations/log.html>: "The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics." Our word vectors are highly skewed distributed, as only a few words in each vector has positive values while others are 0. Thus, log-normalization works well for this problem. We thin TF-IDF may be a better choice for word vector normalization.

(f) The accuracy and confusion matrix for logistic regression are below.

```
Logistic Regression ACC: 0.8150
[[263  37]
 [ 74 226]]
```

The top 20 most important negative words are:

```
Negative top 20:
bad, -2.940010177795963
poor, -2.5000662875487514
worst, -2.1129273163378746
terribl, -1.9187343968442017
wast, -1.8419838593632043
slow, -1.69722945329617
suck, -1.652177338313711
aw, -1.6272464373236337
disappoint, -1.6269823601582507
horribl, -1.4969783123327889
stupid, -1.4604472244941435
start, -1.4338083724143165
bland, -1.4196933573475197
fail, -1.3418801400134233
piec, -1.341395709404752
plot, -1.3377730434028574
rude, -1.3156337005568535
avoid, -1.2928377131075706
hear, -1.2855568225651768
hate, -1.2530753973876092
```

The top 20 most important positive words are:

```
Positive top 20:
great, 3.728217853875945
love, 3.1307304894746175
excel, 2.543477888360755
delici, 2.3579199440102383
nice, 2.2610563814106817
amaz, 2.1328363185656607
fantast, 2.0145059778780205
beauti, 1.9685936229207004
awesom, 1.9095860263657145
best, 1.887332235856134
good, 1.8841164784088518
perfect, 1.7982309244889034
comfort, 1.7010272166626892
wonder, 1.5434490656079216
well, 1.5178364156388924
happi, 1.4414300528592643
incred, 1.4282204915691472
fine, 1.3881251478268803
funni, 1.3345018176751962
sturdi, 1.3205563559302762
```

Naive Bayes accuracy and confusion matrix.

```
Gaussian Naive Bayes ACC: 0.6317
[[268  32]
 [189 111]]
Bernoulli Naive Bayes ACC: 0.8050
[[252  48]
 [ 69 231]]
```

The accuracy and confusion matrix above shows that Logistic Regression and Bernoulli Naive Bayes perform better (ACC is larger than 0.8).

(g) Each feature vectors has 10798 length, i.e. representing 10798 words. The first two reviews' feature vectors are as follows.

```
print training_set.iloc[0]
print training_set.iloc[1]

sentence                                way plug u unless go convert
score                                                                0
feature_vector      [0.69314718056, 0.69314718056, 0.69314718056, ...
2_gram_feature_vector  [1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
Name: 0, dtype: object
sentence                                tie charger convers last minutesmajor problem
score                                                                0
feature_vector      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.69314718056, ...
2_gram_feature_vector  [0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, ...
Name: 3, dtype: object
```

I use log-normalization, too, with the same reason above.

The accuracy and confusion matrix for logistic regression are below.

Logistic Regression ACC: 0.6433

```
[[271  29]
 [185 115]]
```

The top 20 most important negative words are:

Negative top 20:

- wast time, -1.6266337663157204
- wast money, -1.2108755181402313
- custom servic, -0.885446698216792
- poor qualiti, -0.8081575997816369
- stay away, -0.80332538188543
- piec junk, -0.7777574898399167
- worst ever, -0.7631486261640384
- bad film, -0.7557103236968924
- realli bad, -0.7313689428304294
- wait wait, -0.7110396622789038
- good way, -0.7076382278339408
- make mistak, -0.7071255012510449
- buy product, -0.7048544951320117
- ever go, -0.6881550537396549
- zero star, -0.6663238958811999
- act bad, -0.6662888020728832
- go back, -0.6625560941137096
- anytim soon, -0.6614309848132359
- look good, -0.6601001979350706
- send back, -0.6530665613641474

The top 20 most important positive words are:

Positive top 20:

- work great, 2.054405466821827
- high recommend, 1.7524930364450328
- one best, 1.4582411255582608
- great phone, 1.289089164994326
- great product, 1.1872160667930691
- food good, 1.0734850744325828
- realli good, 1.0679433092586532
- easi use, 1.009555329850475
- great food, 0.9843961434373667
- reason price, 0.9029096380166195
- food delici, 0.8999181388557159
- good price, 0.8854444828109942
- great servic, 0.8851646225386762
- love place, 0.8686200339604135
- work fine, 0.835652593585165
- pretti good, 0.8017189465143757
- well made, 0.7951033012950273
- great film, 0.7902639845663137
- good product, 0.7767482726210645
- great place, 0.776207877651663

Naive Bayes accuracy and confusion matrix.

Gaussian Naive Bayes ACC: 0.6300

```
[[282  18]
 [204  96]]
```

Bernoulli Naive Bayes ACC: 0.6400

```
[[273  27]
 [189 111]]
```

The accuracy and confusion matrix above shows that Logistic Regression and Naive Bayes have similar performance.

(h) We implement the PCA using basic matrix operations. The results are listed below.

(1) 1-gram

The accuracy and confusion matrix for logistic regression are below (reduce the dimension of features to 10, 50 and 100, respectively.)

Logistic Regression ACC: 0.5867

```
[[255  45]
 [203  97]]
```

Logistic Regression ACC: 0.6983

```
[[253  47]
 [134 166]]
```

Logistic Regression ACC: 0.7333

```
[[250  50]
 [110 190]]
```

The accuracy and confusion matrix for Naive Bayes are below (reduce the dimension of features to 10, 50 and 100, respectively.)

Gaussian Naive Bayes ACC: 0.5650

```
[[188 112]
 [149 151]]
```

Bernoulli Naive Bayes ACC: 0.5650

```
[[188 112]
 [149 151]]
```

Gaussian Naive Bayes ACC: 0.6233

```
[[190 110]
 [116 184]]
```

Bernoulli Naive Bayes ACC: 0.6233

```
[[190 110]
 [116 184]]
```

Gaussian Naive Bayes ACC: 0.6650

```
[[210  90]
 [111 189]]
```

Bernoulli Naive Bayes ACC: 0.6650

```
[[210  90]
 [111 189]]
```

(2) 2-gram The accuracy and confusion matrix for logistic regression are below (reduce the dimension of features to 10, 50 and 100, respectively.)

Logistic Regression ACC: 0.5050

```
[[297   3]
 [294   6]]
```

Logistic Regression ACC: 0.5267

```
[[292   8]
 [276  24]]
```

Logistic Regression ACC: 0.5367

```
[[286  14]
 [264  36]]
```

The accuracy and confusion matrix for Naive Bayes are below (reduce the dimension of features to 10, 50 and 100, respectively.)

Gaussian Naive Bayes ACC: 0.5033

```
[[292   8]
 [290  10]]
```

Bernoulli Naive Bayes ACC: 0.5033

```
[[292   8]
 [290  10]]
```

Gaussian Naive Bayes ACC: 0.5150

```
[[264  36]
 [255  45]]
```

Bernoulli Naive Bayes ACC: 0.5150

```
[[264  36]
 [255  45]]
```

Gaussian Naive Bayes ACC: 0.5283

```
[[255  45]
 [238  62]]
```

Bernoulli Naive Bayes ACC: 0.5283

```
[[255  45]
 [238  62]]
```

The accuracy and confusion matrix above shows that using PCA may lose some accuracy, i.e. more dimension reduce, more loss. But when we reduce the dimension to 100, the performance is good enough. So sometimes we need to make a trade-off between computing resources and performance, to determine whether to use PCA. We can also find that 2-gram still performs worse than 1-gram.

(i) The bag of words performs best in the task, because it does not reduce any information like PCA and sentiment in reviews is more likely to be reflected by single words rather than 2-grams. People prefer to use single word to express their attitude and sentiment in online reviews, because online reviews are usually extremely short and the words in them

express sentiment with high efficiency.

Problem 2. Clustering for Text Analysis..

(a) doc-word

We use Elbow method to find the best k, but the figure does not have an obvious “elbow”.
We choose 8 as k.

The top 10 words in each cluster are as follows.

cluster 0:

- residues
- crystal
- binding
- conserved
- side
- helix
- loop
- chains
- residue
- structural

cluster 1:

- says
- researchers
- fig
- scientists
- year
- just
- get
- people
- last
- usa

cluster 2:

- protein
- gene
- proteins
- cell
- sequence
- genes
- dna
- cells
- amino

sequences

cluster 3:

cells
expression
cell
protein
mice
expressed
antibody
mouse
induced
expressing

cluster 4:

energy
electron
fig
density
shows
temperature
structure
human
measured
constant

cluster 5:

responses
response
neurons
stimuli
visual
significant
stimulus
fig
test
cortex

cluster 6:

fig
mail
shown
reports
observed
function

correspondence
start
analysis
addressed

cluster 7:

values
global
north
estimates
estimate
surface
years
variations
lower
period

The top 10 documents closest to each cluster center are as follows.

cluster 0:

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
Structure of the S15,S6,S18-rRNA Complex: Assembly of the 30S Ribosome Central Domain
Atomic Structure of PDE4: Insights into Phosphodiesterase Mechanism and Specificity
Twists in Catalysis: Alternating Conformations of Escherichia coli Thioredoxin Reductase
The Productive Conformation of Arachidonic Acid Bound to Prostaglandin Synthase
Redox Signaling in Chloroplasts: Cleavage of Disulfides by an Iron-Sulfur Cluster
Convergent Solutions to Binding at a Protein-Protein Interface
Structural Basis of Smad2 Recognition by the Smad Anchor for Receptor Activation
Structure of the Protease Domain of Memapsin 2 (b-Secretase) Complexed with Inhibitor

cluster 1:

Information Technology Takes a Different Tack
Science Survives in Breakthrough States
Vaccine Studies Stymied by Shortage of Animals
For 'Father' of Abortion Drug, Vindication at Last
On a Slippery Slope to Mediocrity?
In Europe, Hooligans Are Prime Subjects for Research
Japan's 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 2:

- 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
- Ubiquitination: More Than Two to Tango
- Efficient Initiation of HCV RNA Replication in Cell Culture
- New Insights into an Old Modification
- Negative Regulation of the SHATTERPROOF Genes by FRUITFULL during Arabidopsis Fruit Development
- Reading the Worm Genome
- Active Remodeling of Somatic Nuclei in Egg Cytoplasm by the Nucleosomal ATPase ISWI
- Cloning and Heterologous Expression of the Epothilone Gene Cluster

cluster 3:

- T Cell-Independent Rescue of B Lymphocytes from Peripheral Immune Tolerance
- 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
- Patterning of the Zebrafish Retina by a Wave of Sonic Hedgehog Activity
- An Anti-Apoptotic Role for the p53 Family Member, p73, during Developmental Neuron Death
- Impaired Prion Replication in Spleens of Mice Lacking Functional Follicular Dendritic Cells
- Requirement of the RNA Editing Deaminase ADAR1 Gene for Embryonic Erythropoiesis
- CD95/CD95 Ligand Interactions on Epithelial Cells in Host Defense to *Pseudomonas aeruginosa*
- Severely Reduced Female Fertility in CD9-Deficient Mice
- Regulation of B Lymphocyte and Macrophage Development by Graded Expression of PU.1

cluster 4:

- Ambipolar Pentacene Field-Effect Transistors and Inverters
- A Stable Bicyclic Compound with Two Si=Si Double Bonds
- A Cyclic Carbanionic Valence Isomer of a Carbocation: Diphosphino Analogs of Diaminocarocations
- Graphical Evolution of the Arnold Web: From Order to Chaos
- High-Gain Harmonic-Generation Free-Electron Laser
- Prospects for the Polymer Nanoengineer
- Viscosity Mechanisms in Accretion Disks
- Mechanisms of Ordering in Striped Patterns

Anomalous Polarization Profiles in Sunspots: Possible Origin of Umbral Flashes
A Light-Emitting Field-Effect Transistor

cluster 5:

Cholinergic Synaptic Inhibition of Inner Hair Cells in the Neonatal Mammalian Cochlea
Selectivity for 3D Shape That Reveals Distinct Areas within Macaque Inferior Temporal Cortex
Mirror-Image Confusion in Single Neurons of the Macaque Inferotemporal Cortex
Abolition and Reversal of Strain Differences in Behavioral Responses to Drugs of Abuse after a Brief Experience
Reversal of Antipsychotic-Induced Working Memory Deficits by Short-Term Dopamine D1 Receptor Stimulation
Language Discrimination by Human Newborns and by Cotton-Top Tamarin Monkeys
Modulation of Human Visual Cortex by Crossmodal Spatial Attention
Necessity for Afferent Activity to Maintain Eye-Specific Segregation in Ferret Lateral Geniculate Nucleus
Control of SIV Rebound through Structured Treatment Interruptions during Early Infection
Evidence for Brainstem Structures Participating in Oculomotor Integration

cluster 6:

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: Charon's First Detailed Spectra Hold Many Surprises
Corrections and Clarifications: A Short Fe-Fe Distance in Peroxodiferic Ferritin: Control of Fe Substrate versus Cofactor Decay?
Heretical Idea Faces Its Sternest Test
Archaeology in the Holy Land
Corrections and Clarifications: Uninterrupted MCM2-7 Function Required for DNA Replication Fork Progression
Corrections and Clarifications: A Nuclear Solution to Climatic Change?

cluster 7:

Reconstruction of the Amazon Basin Effective Moisture Availability over the past 14,000 Years
Greenland Ice Sheet: High-Elevation Balance and Peripheral Thinning
Isotopic Evidence for Variations in the Marine Calcium Cycle over the Cenozoic
Mass Balance of the Greenland Ice Sheet at High Elevations
Rapid Kimberlite Ascent and the Significance of Ar-Ar Ages in Xenolith Phlogopites
Glacial Climate Instability
The Role of the Southern Ocean in Uptake and Storage of Anthropogenic Carbon Dioxide
Remobilization in the Cratonic Lithosphere Recorded in Polycrystalline Diamond
Accretion of Primitive Planetesimals: Hf-W Isotopic Evidence from Enstatite Chondrites

Temporal Trends in Deep Ocean Redfield Ratios

The algorithm captured the themes of documents, and clustered them into different categories according to their subjects, such as biology, geography and history. This algorithm can be used to classify articles into different categories based on their contents and subjects.

(b) word-doc

We use Elbow method to find the best k. We choose 9 as k.

The top 10 titles in each cluster are as follows.

cluster 0:

- Regulated Cleavage of a Contact-Mediated Axon Repellent
- Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator of p53-Induced Apoptosis
- Positional Syntenic Cloning and Functional Characterization of the Mammalian Circadian Mutation tau
- Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles
- Interacting Molecular Loops in the Mammalian Circadian Clock
- Cross Talk between Interferon-g and -a/b Signaling Components in Caveolar Membrane Domains
- Regulation of STAT3 by Direct Binding to the Rac1 GTPase
- Function of PI3Kg in Thymocyte Development, T Cell Activation, and Neutrophil Migration
- Central Role for G Protein-Coupled Phosphoinositide 3-Kinase g in Inflammation
- Protein Interaction Mapping in C. elegans Using Proteins Involved in Vulval Development

cluster 1:

- Status and Improvements of Coupled General Circulation Models
- Sedimentary Rocks of Early Mars
- Climate Extremes: Observations, Modeling, and Impacts
- A 22,000-Year Record of Monsoonal Precipitation from Northern Chile's Atacama Desert
- Internal Structure and Early Thermal Evolution of Mars from Mars Global Surveyor Topography and Gravity
- Coherent High- and Low-Latitude Climate Variability during the Holocene Warm Period
- Rapid Changes in the Hydrologic Cycle of the Tropical Atlantic during the Last Glacial
- Climate Impact of Late Quaternary Equatorial Pacific Sea Surface Temperature Variations
- The Global Carbon Cycle: A Test of Our Knowledge of Earth as a System
- Causes of Climate Change over the past 1000 Years

cluster 2:

- Advances in the Physics of High-Temperature Superconductivity
- Quantum Criticality: Competing Ground States in Low Dimensions
- The Atom-Cavity Microscope: Single Atoms Bound in Orbit by Single Photons

Orbital Physics in Transition-Metal Oxides
 Negative Poisson's Ratios for Extreme States of Matter
 Generating Solitons by Phase Engineering of a Bose-Einstein Condensate
 Self-Mode-Locking of Quantum Cascade Lasers with Giant Ultrafast Optical Nonlinearities
 NEAR at Eros: Imaging and Spectral Results
 Blue-Fluorescent Antibodies
 The Galactic Center: An Interacting System of Unusual Sources

cluster 3:

Positional Syntenic Cloning and Functional Characterization of the Mammalian Circadian Mutation tau
 The Genome Sequence of *Drosophila melanogaster*
 Kinesin Superfamily Motor Protein KIF17 and mLin-10 in NMDA Receptor-Containing Vesicle Transport
 Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator of p53-Induced Apoptosis
 Crystal Structure of the Ribonucleoprotein Core of the Signal Recognition Particle
 Pol k: A DNA Polymerase Required for Sister Chromatid Cohesion
 Integration of Multiple Signals through Cooperative Regulation of the N-WASP-Arp2/3 Complex
 Role of the Mouse ank Gene in Control of Tissue Calcification and Arthritis
 Regulated Cleavage of a Contact-Mediated Axon Repellent
 Comparative Genomics of the Eukaryotes

cluster 4:

A Mouse Chronology
 Meltdown on Long Island
 Presidential Forum: Gore and Bush Offer Their Views on Science
 Silent No Longer: 'Model Minority' Mobilizes
 Atom-Scale Research Gets Real
 The Rise of the Mouse, Biomedicine's Model Mammal
 Ecologists on a Mission to Save the World
 Infectious History
 Help Needed to Rebuild Science in Yugoslavia
 I'd like to See America Used as a Global Lab

cluster 5:

NEAR at Eros: Imaging and Spectral Results
 Reduction of Tropical Cloudiness by Soot
 The Atom-Cavity Microscope: Single Atoms Bound in Orbit by Single Photons
 Rocks from the Mantle Transition Zone: Majorite-Bearing Xenoliths from Malaita, Southwest Pacific
 Climate Impact of Late Quaternary Equatorial Pacific Sea Surface Temperature Variations

Climate Extremes: Observations, Modeling, and Impacts
 Causes of Climate Change over the past 1000 Years
 Experiments and Simulations of Ion-Enhanced Interfacial Chemistry on Aqueous NaCl
 Aerosols
 Advances in the Physics of High-Temperature Superconductivity
 Internal Structure and Early Thermal Evolution of Mars from Mars Global Surveyor
 Topography and Gravity

cluster 6:

A Mouse Chronology
 Atom-Scale Research Gets Real
 The Genome Sequence of *Drosophila melanogaster*
 Breakthrough of the Year: Genomics Comes of Age
 Presidential Forum: Gore and Bush Offer Their Views on Science
 Infectious History
 Positional Syntenic Cloning and Functional Characterization of the Mammalian Circadian Mutation tau
 Comparative Genomics of the Eukaryotes
 Status and Improvements of Coupled General Circulation Models
 Meltdown on Long Island

cluster 7:

Inhibition of Experimental Liver Cirrhosis in Mice by Telomerase Gene Delivery
 Translating Stem and Progenitor Cell Biology to the Clinic: Barriers and Opportunities
 Function of PI3K α in Thymocyte Development, T Cell Activation, and Neutrophil Migration
 An Oral Vaccine against NMDAR1 with Efficacy in Experimental Stroke and Epilepsy
 Central Role for G Protein-Coupled Phosphoinositide 3-Kinase γ in Inflammation
 Requirement for ROR γ in Thymocyte Survival and Lymphoid Organ Development
 Therapeutic Approaches to Bone Diseases
 Mammalian Neural Stem Cells
 Prostaglandin D_2 as a Mediator of Allergic Asthma
 Bone Resorption by Osteoclasts

cluster 8:

Help Needed to Rebuild Science in Yugoslavia
 Atom-Scale Research Gets Real
 A Mouse Chronology
 Meltdown on Long Island
 Silent No Longer: 'Model Minority' Mobilizes
 I'd like to See America Used as a Global Lab
 Clones: A Hard Act to Follow
 Ecologists on a Mission to Save the World
 Soft Money's Hard Realities

Lee's Special Status Fuels Academy's Rising Reputation

The top 10 words closest to each cluster center are as follows:

cluster 0:

- kinase
- promoter
- polymerase
- staining
- pcr
- mrna
- vivo
- regulated
- assay
- signaling

cluster 1:

- decadal
- holocene
- anomaly
- sst
- basins
- tropics
- equator
- sedimentary
- lunar
- silicate

cluster 2:

- resonant
- anisotropic
- fermi
- crystallographic
- reflections
- tunneling
- metallic
- lying
- transverse
- incident

cluster 3:

- triton
- methionine
- glutathione

agarose
isoforms
glycerol
histidine
cys
subcellular
glycine

cluster 4:

celera
intelligence
managers
income
schools
math
weapons
capital
court
draft

cluster 5:

start
gray
decrease
error
peak
res
magnitude
fraction
rev
maximum

cluster 6:

aptamers
lcts
dnag
trxr
neas
doxy
proteorhodopsin
lg268
nompc
rory

cluster 7:

lymphoid
immunoreactivity
injury
transplantation
cd8
abnormalities
littermates
hematopoietic
systemic
inflammatory

cluster 8:

researcher
didnt
doesnt
hopes
got
plans
getting
biologist
cant
theres

The algorithm captured the meanings of words, and clustered them into different categories according to the subjects the words belong to, such as biology, geography and history. This algorithm can be used to classify words into different categories based on their meanings.