Results

The homework instructions said to include our data, so I assumed you wanted to see our results. I copied the results from running the code for each section below.

Q.1

total number of reviews: 30000

total number of positive reviews: 15091

total number of negative reviews: 14909

the 119835

and 64619

i 63045

a 60750

to 57968

it 47813

of 47382

this 44363

is 41185

in 27962

number of unique words: 69034

total number of word occurrences: 2384092

('of', 'the') 15208

('the', 'of') 12728

('to', 'the') 11803

('this', 'is') 10832

('the', 'the') 10615

('and', 'the') 10479

('in', 'the') 9092

('the', 'and') 8728

('the', 'is') 8420

('is', 'a') 8106

('blah', 'blah') 14.354618369750598

('sci', 'fi') 13.988916953831932

('hip', 'hop') 13.951136560300908

('harry', 'potter') 13.888268715726547

('stainless', 'steel') 13.603000944936795

('blu', 'ray') 12.876532236161433

('buyer', 'beware') 12.535026185951509

('windows', 'xp') 12.19632379175091

('tech', 'support') 11.499355382516233

('web', 'site') 11.497319189050762

('yourself', 'favor') 11.443661882245157

('page', 'turner') 11.400172770446485

('customer', 'service') 11.278462830751023

('science', 'fiction') 11.096632937974082

('e', 'g') 11.035449516180822

('greatest', 'hits') 11.024506730967389

('special', 'effects') 10.993018359099938

('step', 'step') 10.966224236101938

('coffee', 'maker') 10.914604488783175

('mp3', 'player') 10.865923489098625

('star', 'wars') 10.813039701136118

('u', 's') 10.760542221923275

('stopped', 'working') 10.587777480732095

('fell', 'apart') 10.551845050793398

('common', 'sense') 10.354843984097576

('rock', 'roll') 10.343880980164768

('character', 'development') 10.323053577169755

('thought', 'provoking') 10.271983096742455

('needless', 'say') 10.207728555023158

('local', 'store') 9.926442444633143

('piece', 'junk') 9.910263183811226

('star', 'rating') 9.807004459229624

('highly', 'recommended') 9.746038914681082

('main', 'character') 9.733242590663572

('amazon', 'com') 9.724501019112834

('subject', 'matter') 9.715448056811853

('zero', 'stars') 9.650193218328079

('new', 'york') 9.60604949151238

('yr', 'old') 9.598002274916912

('cheaply', 'made') 9.556727292739176

('15', 'minutes') 9.521283611243884

('look', 'elsewhere') 9.503182793861079

('bottom', 'line') 9.492884092031774

('black', 'white') 9.422480828255283

('high', 'school') 9.39957738022105

('rip', 'off') 9.37143686832167

('stay', 'away') 9.280159634779801

('looking', 'forward') 9.185879686680382

('30', 'minutes') 9.143197248137932

('20', 'minutes') 9.062931974377614

('years', 'ago') 9.04496775702666

('highly', 'recommend') 8.998428888328023

('five', 'stars') 8.933328830218382

('point', 'view') 8.932903303911111

('positive', 'reviews') 8.921246930647163

('start', 'finish') 8.902958581596154

('6', 'months') 8.901769702556953

('waste', 'money') 8.877840397996444

('ripped', 'off') 8.848120257239271

('who', 'cares') 8.840349808294468

('light', 'weight') 8.821395884427215

('save', 'money') 8.807879698453423

('huge', 'fan') 8.801151604984748

('daughter', 'loves') 8.783446970662089

('die', 'hard') 8.721654405085756

('save', 'yourself') 8.715566822460529

('made', 'china') 8.702805552227119

('poorly', 'written') 8.692690671272684

("can't", 'wait') 8.682695211749477

('5', 'stars') 8.640547080990537

('easy', 'install') 8.601868713408736

('insight', 'into') 8.601768973006962

('year', 'old') 8.552084235441175

('son', 'loves') 8.52763748227387

('10', 'minutes') 8.460369574515184

('world', 'war') 8.442776452949666

('wake', 'up') 8.417454564523446

('christmas', 'gift') 8.413490050000407

('must', 'admit') 8.411869271803385

('picked', 'up') 8.411468269457714

('total', 'waste') 8.4003914125171

('ended', 'up') 8.37616802857631

("haven't", 'seen') 8.367658928078624

('other', 'reviewers') 8.362686945390557

('hard', 'drive') 8.357245821785382

('anyone', 'interested') 8.328097731991239

('pick', 'up') 8.300985053893685

('do', 'favor') 8.279117254302813

('look', 'forward') 8.249537950217647

('grew', 'up') 8.2429760767943

('everyone', 'else') 8.226580558817547

('months', 'ago') 8.221402877915612

('weeks', 'ago') 8.20136705477172

('writing', 'style') 8.194157690231863

('put', 'together') 8.186661941914721

("can't", 'imagine') 8.184092995831502

('anything', 'else') 8.129360260468774

('someone', 'else') 8.125990008185507

("i've", 'seen') 8.12095379653595

('another', 'reviewer') 8.11426635210301

Q.2

QUESTION 2

the 13245

and 12526

a 11891

to 11156

this 11082

i 10334

is 10106

of 9998

it 9793

in 8182

the 13615

and 11743

a 11675

to 11463

i 11415

this 11379

it 10358

of 10128

is 9420

not 8236

great 2259.208635662306

not 2016.4733034985088

waste 1300.8223017245875

money 1204.9485759018141

love 769.4901479541076

best 757.4894506044782

poor 667.987842854487

worst 663.642219565316

excellent 646.8180379179539

disappointed 636.8233634415676

bad 592.3330800335547

easy 474.1369365588785

terrible 471.9743085383744

was 451.91998688865056

boring 450.7034678808181

disappointing 447.68946599350164

wonderful 442.9011326531304

don't 438.4562426890645

horrible 395.1435193557931

highly 385.7808060602993

no 379.0378467783738

perfect 349.0765845868453

return 344.6818564536531

amazing 332.9478002391661

disappointment 330.6399879586602

awful 329.0240374012976

nothing 324.77106972343216

would 320.6329630624985

favorite 312.22858840527005

poorly 300.49670839651225

junk 290.2668010951129

awesome 284.75936622066524

useless 273.95049348683244

fun 271.8741573291675

buy 267.7404278957863

after 263.8238682356786

didn't 255.7262442466906

returned 250.98991415659458

save 246.49520805685768

i 245.96471630382592

tried 244.2432935462157

garbage 240.6227027722259

instead 238.43162068558706

do 233.5575929321808

must 233.3903210185869

loved 226.11170640604948

loves 225.5019559487964

fantastic 224.55798651625818

did 218.5434397248459

beautiful 218.40451110949002

unfortunately 216.62572031667986

but 205.4477958413108

well 204.73401371222843

worse 204.72280389745723

refund 201.23174116409194

even 194.1072549104398

product 189.5780181784276

doesn't 188.8539581133838

classic 183.17693038098315

beware 177.3305445900168

off 169.4937286278893

trash 168.28684616488644

same 167.9499653932283

if 166.7651347646435

broke 166.03242079894147

cheap 165.4715327880216

enjoyed 165.09701906818972

stopped 163.30561697503953

reviews 160.70180654227275

sorry 160.5485056834657

maybe 159.8322403661031

wasted 158.56500517823287

stupid 156.24547975165058

had 154.2115384658771

pay 152.10330790082818

then 151.20946627927145

returning 146.33251374570366

be 143.28220259848797

trying 142.65833936941115

wouldn't 141.3670953977469

bought 140.56667967233483

worthless 138.8826043681111

bother 138.8537232025012

avoid 138.03126754325513

defective 135.73122410925868

please 135.64884524390334

ok 134.5671532557336

customer 134.34079989426047

weak 133.49482489759387

sent 132.398607367683

good 131.14661489000534

were 130.94015823138227

ridiculous 129.58692111581254

guess 128.66998202961378

sucks 128.60204643691716

pleased 126.91687574588896

too 126.59769442978774

buying 126.35924582586985

enjoyable 126.15070955743849

unless 125.8048313994943

jennagatley@Jennas-MBP-4 ~ %

Q.3

number of words/phrases in the list: 305868

Input String: abreviation

Top 10 similar words using Jaccard similarity (tri-grams):

[('abbreviation', 0.7272727272727273), ('abbreviations', 0.6666666666666666), ('abbreviationi', 0.6666666666666666), ('abbreviatione', 0.6666666666666666), ('adbreviationi', 0.6666666666666666), ('adbreviatione', 0.6666666666666666), ('abbreviatio', 0.6363636363636364), ('adbreviatio', 0.6363636363636364), ('abreviativo', 0.6363636363636364), ('abreviativa', 0.6363636363636364)]

Top 10 similar words using edit distance:

[('abbreviation', 1), ('abbreviatio', 2), ('abbreviations', 2), ('alleviation', 2), ('abbreviationi', 2), ('abbreviatione', 2), ('abreviaron', 2), ('adbreviatio', 2), ('adbreviationi', 2), ('adbreviatione', 2)]

Input String: abstrictiveness

Top 10 similar words using Jaccard similarity (tri-grams):

[('abstractiveness', 0.625), ('activeness', 0.5), ('addictiveness', 0.5), ('astrictive', 0.5), ('abstriction', 0.4666666666666667), ('abstricting', 0.4666666666666667), ('astrictives', 0.4666666666666667), ('abstrict', 0.46153846153846156), ('abstrictions', 0.4375), ('activenesses', 0.4375)]

Top 10 similar words using edit distance:

[('abstractiveness', 1), ('absorptiveness', 3), ('attractiveness', 3), ('abortiveness', 4), ('abstersiveness', 4), ('abstractedness', 4), ('abstractness', 4), ('assertiveness', 4), ('abstrictions', 4), ('attributiveness', 4)]

Input String: accanthopterigious

Top 10 similar words using Jaccard similarity (tri-grams):

[('acanthopterygious', 0.55), ('acanthopteri', 0.5294117647058824), ('acanthopteri', 0.5294117647058824), ('acanthopterous', 0.47368421052631576), ('acanthopteran', 0.42105263157894735), ('acanthopterans', 0.4), ('acanthopterygii', 0.38095238095238093), ('acanthopterygian', 0.36363636363636365), ('acanthopt', 0.35294117647058826), ('acanthopterygians', 0.34782608695652173)]

Top 10 similar words using edit distance:

[('acanthopterygious', 2), ('acanthopterous', 4), ('acanthopterygians', 4), ('acanthopterygian', 5), ('acanthopterygii', 5), ('acanthopodious', 6), ('acanthophorous', 6), ('acanthopteri', 6), ('acanthopteri', 6), ('acanthopterans', 6)]

Input String: artifitial inteligwnse

Top 10 similar words using Jaccard similarity (tri-grams):

[('artificial intelligence', 0.41379310344827586), ('artificial intelligences', 0.4), ('artificial insemination', 0.28125), ('artificial art', 0.24), ('artificial anus', 0.2222222222222222), ('artificial life', 0.2222222222222222), ('artificialmente', 0.2222222222222222), ('artificialities', 0.2222222222222222), ('artificial turf', 0.2222222222222222), ('artificial', 0.21739130434782608)]

Top 10 similar words using edit distance:

[('artificial intelligence', 4), ('artificial intelligences', 5), ('artificial life', 9), ('artificial insemination', 9), ('artificialities', 9), ('artificial horizons', 9), ('artificial persons', 9), ('artificial abortions', 9), ('artificial language', 10), ('adaptive intelligence', 10)]

Input String: agglumetation

Top 10 similar words using Jaccard similarity (tri-grams):

[('agglutination', 0.375), ('arietation', 0.35714285714285715), ('agglutinations', 0.35294117647058826), ('arietationi', 0.3333333333333333), ('arietatione', 0.3333333333333333), ('arietations', 0.3333333333333333), ('arietationes', 0.3125), ('arietationis', 0.3125), ('arietationum', 0.3125), ('arietationem', 0.3125)]

Top 10 similar words using edit distance:

[('agglomeration', 2), ('agglomeration', 2), ('agglutination', 3), ('agglomerations', 3), ('argumentation', 3), ('argumentation', 3), ('agglomeratioun', 3), ('acclimatation', 4), ('augmentation', 4), ('aggregation', 4)]

BIGRAMS TEST

Input String: abreviation

Top 10 similar words using Jaccard similarity (Bi-grams):

[('abbreviation', 0.9090909090909091), ('abbreviations', 0.8333333333333334), ('abbreviationi', 0.8333333333333334), ('abbreviatione', 0.8333333333333334), ('abbreviatio', 0.8181818181818182), ('abbreviationis', 0.7692307692307693), ('abbreviationem', 0.7692307692307693), ('abbreviationes', 0.7692307692307693), ('abbreviationum', 0.7692307692307693), ('abbreviati', 0.7272727272727273)]

Top 10 similar words using edit distance:

[('abbreviation', 1), ('abbreviatio', 2), ('abbreviations', 2), ('alleviation', 2), ('abbreviationi', 2), ('abbreviatione', 2), ('abreviaron', 2), ('adbreviatio', 2), ('adbreviationi', 2), ('adbreviatione', 2)]

Input String: abstrictiveness

Top 10 similar words using Jaccard similarity (Bi-grams):

[('abstractiveness', 0.75), ('astrictives', 0.6), ('activeness', 0.5333333333333333), ('abstivesen', 0.5333333333333333), ('astrictive', 0.5333333333333333), ('abstivesse', 0.5333333333333333), ('abstivesses', 0.5333333333333333), ('addictiveness', 0.5294117647058824), ('absorptiveness', 0.5), ('abstersiveness', 0.5)]

Top 10 similar words using edit distance:

[('abstractiveness', 1), ('absorptiveness', 3), ('attractiveness', 3), ('abortiveness', 4), ('abstersiveness', 4), ('abstractedness', 4), ('abstractness', 4), ('assertiveness', 4), ('abstrictions', 4), ('attributiveness', 4)]

Input String: accanthopterigious

Top 10 similar words using Jaccard similarity (Bi-grams):

[('acanthopterygious', 0.7368421052631579), ('acanthopterous', 0.6666666666666666), ('acanthopteri', 0.6470588235294118), ('acanthopteri', 0.6470588235294118), ('acanthopteran', 0.5555555555555556), ('acanthopterygian', 0.55), ('acanthopterygii', 0.55), ('acanthopterans', 0.5263157894736842), ('acanthopterygians', 0.5238095238095238), ('acanthopodious', 0.5)]

Top 10 similar words using edit distance:

[('acanthopterygious', 2), ('acanthopterous', 4), ('acanthopterygians', 4), ('acanthopterygian', 5), ('acanthopterygii', 5), ('acanthopodious', 6), ('acanthophorous', 6), ('acanthopteri', 6), ('acanthopteri', 6), ('acanthopterans', 6)]

Input String: artifitial inteligwnse

Top 10 similar words using Jaccard similarity (Bi-grams):

[('artificial intelligence', 0.5555555555555556), ('artificial intelligences', 0.5357142857142857), ('artificial insemination', 0.42857142857142855), ('artificiality', 0.391304347826087), ('artificial life', 0.375), ('artificialities', 0.375), ('artificiali', 0.36363636363636365), ('artigliante', 0.36363636363636365), ('artificialia', 0.36363636363636365), ('artificialmente', 0.36)]

Top 10 similar words using edit distance:

[('artificial intelligence', 4), ('artificial intelligences', 5), ('artificial life', 9), ('artificial insemination', 9), ('artificialities', 9), ('artificial horizons', 9), ('artificial persons', 9), ('artificial abortions', 9), ('artificial language', 10), ('adaptive intelligence', 10)]

Input String: agglumetation

Top 10 similar words using Jaccard similarity (Bi-grams):

[('agglutination', 0.5333333333333333), ('agglomeration', 0.5), ('agglutinations', 0.5), ('agglomeration', 0.5), ('agglomerations', 0.47058823529411764), ('agglutinatarum', 0.47058823529411764), ('agglutinata', 0.4666666666666667), ('agitationum', 0.4666666666666667), ('agglomerationen', 0.4444444444444444), ('autoagglutination', 0.4444444444444444)]

Top 10 similar words using edit distance:

[('agglomeration', 2), ('agglomeration', 2), ('agglutination', 3), ('agglomerations', 3), ('argumentation', 3), ('argumentation', 3), ('agglomeratioun', 3), ('acclimatation', 4), ('augmentation', 4), ('aggregation', 4)]

4-GRAMS TEST

Input String: abreviation

Top 10 similar words using Jaccard similarity (4-grams):

[('abbreviation', 0.7), ('abbreviations', 0.6363636363636364), ('abbreviationi', 0.6363636363636364), ('abbreviatione', 0.6363636363636364), ('adbreviationi', 0.6363636363636364), ('adbreviatione', 0.6363636363636364), ('abbreviatio', 0.6), ('adbreviatio', 0.6), ('abreviativo', 0.6), ('abreviativa', 0.6)]

Top 10 similar words using edit distance:

[('abbreviation', 1), ('abbreviatio', 2), ('abbreviations', 2), ('alleviation', 2), ('abbreviationi', 2), ('abbreviatione', 2), ('abreviaron', 2), ('adbreviatio', 2), ('adbreviationi', 2), ('adbreviatione', 2)]

Input String: abstrictiveness

Top 10 similar words using Jaccard similarity (4-grams):

[('abstractiveness', 0.5), ('addictiveness', 0.4666666666666667), ('activeness', 0.46153846153846156), ('astrictive', 0.46153846153846156), ('abstriction', 0.42857142857142855), ('abstricting', 0.42857142857142855), ('astrictives', 0.42857142857142855), ('abstrict', 0.4166666666666667), ('abstrictions', 0.4), ('activenesses', 0.4)]

Top 10 similar words using edit distance:

[('abstractiveness', 1), ('absorptiveness', 3), ('attractiveness', 3), ('abortiveness', 4), ('abstersiveness', 4), ('abstractedness', 4), ('abstractness', 4), ('assertiveness', 4), ('abstrictions', 4), ('attributiveness', 4)]

Input String: accanthopterigious

Top 10 similar words using Jaccard similarity (4-grams):

[('acanthopteri', 0.5), ('acanthopteri', 0.5), ('acanthopterygious', 0.45), ('acanthopteran', 0.3888888888888889), ('acanthopterous', 0.3684210526315789), ('acanthopterans', 0.3684210526315789), ('acanthopterygii', 0.35), ('acanthopterygian', 0.3333333333333333), ('acanthopterygians', 0.3181818181818182), ('acanthopt', 0.3125)]

Top 10 similar words using edit distance:

[('acanthopterygious', 2), ('acanthopterous', 4), ('acanthopterygians', 4), ('acanthopterygian', 5), ('acanthopterygii', 5), ('acanthopodious', 6), ('acanthophorous', 6), ('acanthopteri', 6), ('acanthopteri', 6), ('acanthopterans', 6)]

Input String: artifitial inteligwnse

Top 10 similar words using Jaccard similarity (4-grams):

[('artificial intelligence', 0.3), ('artificial intelligences', 0.2903225806451613), ('artificial insemination', 0.18181818181818182), ('artificial art', 0.15384615384615385), ('artificial anus', 0.14814814814814814), ('artificial life', 0.14814814814814814), ('artificial turf', 0.14814814814814814), ('artifiko', 0.14285714285714285), ('artifice', 0.14285714285714285), ('artifici', 0.14285714285714285)]

Top 10 similar words using edit distance:

[('artificial intelligence', 4), ('artificial intelligences', 5), ('artificial life', 9), ('artificial insemination', 9), ('artificialities', 9), ('artificial horizons', 9), ('artificial persons', 9), ('artificial abortions', 9), ('artificial language', 10), ('adaptive intelligence', 10)]

Input String: agglumetation

Top 10 similar words using Jaccard similarity (4-grams):

[('arietation', 0.3076923076923077), ('arietationi', 0.2857142857142857), ('arietatione', 0.2857142857142857), ('arietations', 0.2857142857142857), ('arietationes', 0.26666666666666666), ('arietationis', 0.26666666666666666), ('arietationum', 0.26666666666666666), ('arietationem', 0.26666666666666666), ('agglutination', 0.25), ('agglutinations', 0.23529411764705882)]

Top 10 similar words using edit distance:

[('agglomeration', 2), ('agglomeration', 2), ('agglutination', 3), ('agglomerations', 3), ('argumentation', 3), ('argumentation', 3), ('agglomeratioun', 3), ('acclimatation', 4), ('augmentation', 4), ('aggregation', 4)]

5-GRAMS TEST

Input String: abreviation

Top 10 similar words using Jaccard similarity (5-grams):

[('abbreviation', 0.6666666666666666), ('abbreviations', 0.6), ('abbreviationi', 0.6), ('abbreviatione', 0.6), ('adbreviationi', 0.6), ('adbreviatione', 0.6), ('abbreviatio', 0.5555555555555556), ('adbreviatio', 0.5555555555555556), ('abreviativo', 0.5555555555555556), ('abreviativa', 0.5555555555555556)]

Top 10 similar words using edit distance:

[('abbreviation', 1), ('abbreviatio', 2), ('abbreviations', 2), ('alleviation', 2), ('abbreviationi', 2), ('abbreviatione', 2), ('abreviaron', 2), ('adbreviatio', 2), ('adbreviationi', 2), ('adbreviatione', 2)]

Input String: abstrictiveness

Top 10 similar words using Jaccard similarity (5-grams):

[('addictiveness', 0.42857142857142855), ('activeness', 0.4166666666666667), ('astrictive', 0.4166666666666667), ('abstriction', 0.38461538461538464), ('abstricting', 0.38461538461538464), ('astrictives', 0.38461538461538464), ('abstractiveness', 0.375), ('abstrict', 0.36363636363636365), ('abstrictions', 0.35714285714285715), ('activenesses', 0.35714285714285715)]

Top 10 similar words using edit distance:

[('abstractiveness', 1), ('absorptiveness', 3), ('attractiveness', 3), ('abortiveness', 4), ('abstersiveness', 4), ('abstractedness', 4), ('abstractness', 4), ('assertiveness', 4), ('abstrictions', 4), ('attributiveness', 4)]

Input String: accanthopterigious

Top 10 similar words using Jaccard similarity (5-grams):

[('acanthopteri', 0.4666666666666667), ('acanthopteri', 0.4666666666666667), ('acanthopteran', 0.35294117647058826), ('acanthopterygious', 0.35), ('acanthopterous', 0.3333333333333333), ('acanthopterans', 0.3333333333333333), ('acanthopterygii', 0.3157894736842105), ('acanthopterygian', 0.3), ('acanthopterygians', 0.2857142857142857), ('acanthopt', 0.26666666666666666)]

Top 10 similar words using edit distance:

[('acanthopterygious', 2), ('acanthopterous', 4), ('acanthopterygians', 4), ('acanthopterygian', 5), ('acanthopterygii', 5), ('acanthopodious', 6), ('acanthophorous', 6), ('acanthopteri', 6), ('acanthopteri', 6), ('acanthopterans', 6)]

Input String: artifitial inteligwnse

Top 10 similar words using Jaccard similarity (5-grams):

[('artificial intelligence', 0.23333333333333334), ('artificial intelligences', 0.22580645161290322), ('artificial insemination', 0.12121212121212122), ('artifiko', 0.1), ('artifice', 0.1), ('artifici', 0.1), ('artifizi', 0.1), ('artificer', 0.09523809523809523), ('artifical', 0.09523809523809523), ('artifices', 0.09523809523809523)]

Top 10 similar words using edit distance:

[('artificial intelligence', 4), ('artificial intelligences', 5), ('artificial life', 9), ('artificial insemination', 9), ('artificialities', 9), ('artificial horizons', 9), ('artificial persons', 9), ('artificial abortions', 9), ('artificial language', 10), ('adaptive intelligence', 10)]

Input String: agglumetation

Top 10 similar words using Jaccard similarity (5-grams):

[('arietation', 0.25), ('arietationi', 0.23076923076923078), ('arietatione', 0.23076923076923078), ('arietations', 0.23076923076923078), ('arietationes', 0.21428571428571427), ('arietationis', 0.21428571428571427), ('arietationum', 0.21428571428571427), ('arietationem', 0.21428571428571427), ('arietationibus', 0.1875), ('agitation', 0.16666666666666666)]

Top 10 similar words using edit distance:

[('agglomeration', 2), ('agglomeration', 2), ('agglutination', 3), ('agglomerations', 3), ('argumentation', 3), ('argumentation', 3), ('agglomeratioun', 3), ('acclimatation', 4), ('augmentation', 4), ('aggregation', 4)]

jennagatley@Jennas-MBP-4 ~ %