Project 4 Report - CMSC 409 - Artificial Intelligence

Steven Hernandez

In total, there are 196 unique root words found.

24 words that are encountered at least 2 times.

And then only 18 words that are encountered at least 3 times.

These statistics are calculated based on processing the documents in the following ways:

a. Tokenizing the sentences, which splits each sentence on the spaces to only produce a list of word/numeric tokens. This allows us to begin processing each word individually without requiring the context of the entire sentence.
b. Removing punctuation is required because in general, punctuation does not provide us textual context. Again, we are only looking at the similarity of sentence based on the number of occurrences of common words between the sentences. We are not trying to decifer the intent or the sentiment behind the sentence, so we do not require punctuation or even placement of words within the sentence. Just that the word exists within the sentence.
c. Removing numbers because numbers do not provide context about what the sentence is talking about. A book might cost $20 as would a basic microcontroller like an Arduino, but they are not related. Additional since, we removed punctuation in the previous step, we wouldn't be able to differentiate $20 from 20 miles or 20 participants, etc.
d. Converting upper to lower case prevents words appearing at the beginning of a sentence (with a required capital letter) from being considered a different word if it also appears in the middle of a sentence (which would be written in all lower case)
e. Removing stop words shrinks the total number of words that we find. More importantly though, it removes overly common words that do not provide us useful insights into the similarity of sentences. The word 'the' is very likely to appear in most sentences, thus is not a useful indicator.
f. Stemming takes a word in past tense/future tense or plural/singular and takes the 'stem' or 'root' word. This further shrinks the overall number of words or dimensions that we must analyze. An example: run and running have the same root word, thus are very similar.
g. Combining stemmed words takes these common stemmed root words and combines them so that we can get a total count of the occurances of the word throughout all sentence documents.

On the following page is a table listing all of these root words along with the number of occurrences of the word through the documents (the feature vector)

Root Word	# of instances	Root Word	# of instances	Root Word	# of instances	Root Word	# of instances
autonom	5	musk	1	public	1	reason	1
sedan	3	wai	2	us	1	attain	1
travel	1	escap	1	driven	1	multipl	1
type	1	obsolesc	1	combin	1	domain	1
road	5	have	1	two	2	feel	1
speed	1	sort	1	paradigm	1	express	1
up	2	merger	1	everyth	1	understand	2
mile	10	biolog	1	know	1	emot	1
per	7	entir	1	realiti	1	singl	1
hour	6	interior	1	sens	2	famili	1
futur	1	freshli	1	knowledg	2	conveni	1
machin	6	paint	1	experi	1	locat	1
learn	5	go	1	via	1	near	1
rai	2	percent	1	five	1	major	1
kurzweil	2	befor	1	on	2	rout	1
predict	1	lap	2	suit	1	author	1
year	1	possibl	2	king	1	book	1
awai	1	lead	2	bed	1	ag	1
singular	1	necessarili	1	veri	2	spiritu	1
selfimprov	1	sentienc	1	nice	1	describ	1
artifici	6	applianc	2	queen	1	spread	1
superintellig	1	includ	2	complet	1	throughout	1
far	1	well	1	over	2	cosmo	1
exce	1	secur	1	drive	2	cute	1
human	2	system	1	accidentfre	1	classi	1
intellig	11	tenant	1	updat	3	open	1
get	1	respons	1	renov	1	area	1
car	5	electr	1	new	1	great	1
kilomet	6	water	2	floor	1	went	1
second	2	ga	1	john	1	round	1
newli	1	pet	2	mccarthi	1	minut	1
remodel	1	negoti	1	inventor	1	recent	1
home	4	base	1	program	2	work	1
rent	1	anim	1	languag	1	fundament	1
bedroom	7	four	1	lisp	1	techniqu	1
bath	3	row	1	coin	1	deep	1
live	4	hous	3	term	1	lai	1
room	4	come	3	deal	1	groundwork	1
larg	3	washer	1	number	1	automat	1
eat	2	dryer	1	situat	1	through	1
kitchen	5	finish	1	averag	2	increas	1
full	1	basement	1	gallon	1	world	1
size	4	three	1	pound	1	bathroom	1
util	1	park	1	gener	1	townhous	1
test	2	space	1	t	1	central	1
achiev	1	back	1	selfawar	1	heat	1
rang	1	approv	1	comput	2	air	1
around	4	owner	1	engag	1	trash	1
charg	2	limit	1	commonsens	1	sewag	1

The following 2 tables show the distribution of root words which appear at least 2 times across each document (with each row indicating one sentence) (This is the Term Document Matrix **TDM**)

autonom	sedan	road	mile	per	hour	machin	learn	artifici	intellig	car	kilomet
1	1	1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1	0	0
0	0	0	1	2	2	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	0	0	2	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	1	1
1	1	0	1	2	2	0	0	0	0	0	1
0	0	0	0	0	0	2	1	1	3	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1	0	0
1	0	1	0	0	0	0	0	0	0	1	0
0	1	0	2	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	2	1	1	0	0	0	0	1	0
0	0	0	0	0	0	1	2	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0

bedroom	bath	live	room	larg	kitchen	size	around	hous	come	updat
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
1	1	1	2	1	1	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	1	1	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	1	0	3	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	1	0	2
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	2	0	0	1	0	1
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
1	0	1	1	0	1	0	0	0	0	0