Jacob Wilson

Jacob Roberts

Project 4 Report

1.

Feature Vector is [ 'autonom', 'sedan', 'travel', 'type', 'road', 'speed', 'up', 'mile', 'per', 'hour', 'futur', 'machin', 'learn', 'rai', 'kurzweil', 'predict', 'year', 'awai', 'singular', 'self', 'improv', 'artifici', 'super', 'intellig', 'far', 'exce', 'human', 'get', 'car', 'kilomet', 'second', 'newli', 'remodel', 'home', 'rent', 'bedroom', 'bath', 'live', 'room', 'larg', 'eat', 'kitchen', 'full', 'size', 'util', 'test', 'achiev', 'rang', 'around', 'charg', 'musk', 'wai', 'escap', 'obsolesc', 'have', 'sort', 'merger', 'biolog', 'entir', 'interior', 'freshli', 'paint', 'go', 'percent', 'befor', 'lap', 'possibl', 'lead', 'necessarili', 'sentienc', 'applianc', 'includ', 'well', 'secur', 'system', 'tenant', 'respons', 'electr', 'water', 'ga', 'pet', 'negoti', 'base', 'anim', 'four', 'row', 'hous', 'come', 'washer', 'dryer', 'finish', 'basement', 'three', 'park', 'space', 'back', 'approv', 'owner', 'limit', 'public', 'us', 'driven', 'combin', 'two', 'paradigm', 'everyth', 'know', 'realiti', 'sens', 'knowledg', 'experi', 'via', 'five', 'on', 'suit', 'king', 'bed', 'veri', 'nice', 'queen', 'complet', 'over', 'drive', 'accid', 'free', 'updat', 'renov', 'new', 'floor', 'john', 'mccarthi', 'inventor', 'program', 'languag', 'lisp', 'coin', 'term', 'deal', 'number', 'situat', 'averag', 'gallon', 'pound', 'gener', 't', 'awar', 'comput', 'engag', 'common', 'reason', 'attain', 'multipl', 'domain', 'feel', 'express', 'understand', 'emot', 'singl', 'famili', 'conveni', 'locat', 'near', 'major', 'rout', 'author', 'book', 'ag', 'spiritu', 'describ', 'spread', 'throughout', 'cosmo', 'cute', 'classi', 'open', 'area', 'great', 'went', 'round', 'minut', 'recent', 'work', 'fundament', 'techniqu', 'deep', 'lai', 'groundwork', 'automat', 'through', 'increas', 'world', 'bathroom', 'townhous', 'central', 'heat', 'air', 'trash', 'sewag' ]

2.1

See tdm.xlsx for the Term Document Matrix (NOTE: The program generates the tdm as a csv file)

2.2

A: Tokenize Sentences: This allows us to address each sentence separately and obtain per sentence metrics which is very useful statistically. The drawback to this is contextual information that spans multiple sentences is lost.

B: Remove Punctuation: This allows us to focus on words. Punctuation turns into noise when taken out of context but we do loose some grammatical information of the sentence.

C: Remove Numbers: This also allows us to focus on the words. We do this because the numbers have no meaning when taken out of context but as a result we loose some precision that would be useful in information retrieval.

D: Convert upper to lowercase: This removes extraneous detail that has no useful information out of context. However this information can be useful for proper noun detection.

E: Remove stop words: This removes words that are extremely common in the English language and therefore have no meaning when taken out of context.

F: Perform Stemming: Reduces words down to their base form which carry the most useful information. This allows us to avoid alternate versions of the same words skewing any learning. The information lost here would useful for things such as tense recognition.

G: Combine Stemmed Words: Here we take the stemmed words can combine them back into a psuedo sentence for further processing and analytics among the sentences, specifically to build the term document matrix. No information is lost here.

3.