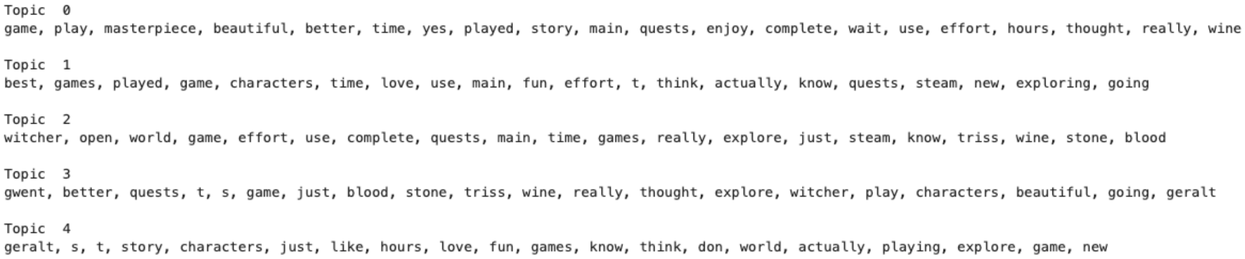
Unsupervised MVP

The goal of this project was to model topics in reviews of popular video games, to create a recommendation system. The idea is that, by recommending games based on topics found in reviews, users will be exposed to new games and genres they would otherwise look over, opposed to sticking to games from a particular genre. To begin this process, and practice NLP skills, I looked at just 20 reviews from The Witcher 3, which is the highest ranked RPG (role playing game) listed on SteamPowered.com (the data source for this project. I tried both CountVectorizer and TF-IDF vectorization, with both LSA and NMF. What appeared to work the best was TF-IDF and NMF. I started with looking at 5 topics. With such a small corpus, not all topics were great ones, but some did have clear ones:



```
Topic  0
game, play, masterpiece, beautiful, better, time, yes, played, story, main, quests, enjoy, complete, wait, use, effort, hours, thought, really, wine

Topic  1
best, games, played, game, characters, time, love, use, main, fun, effort, t, think, actually, know, quests, steam, new, exploring, going

Topic  2
witcher, open, world, game, effort, use, complete, quests, main, time, games, really, explore, just, steam, know, triss, wine, stone, blood

Topic  3
gwent, better, quests, t, s, game, just, blood, stone, triss, wine, really, thought, explore, witcher, play, characters, beautiful, going, geralt

Topic  4
geralt, s, t, story, characters, just, like, hours, love, fun, games, know, think, don, world, actually, playing, explore, game, new
```

The first topic seem to relate to gameplay, in the sense of a good story, good quests that take time, while the last one seems to most strongly relate to characters. Please note as this was the first iteration, no lemmatization was applied. Passing a sample query of “open world” through these topics showed a strong correlation between topic 2, and a slight correlation to topic 4, which goes along with the words we can see printed on the screen, and pairwise_distances, I was able to look at which reviews most strongly displayed those topics.

Since this iteration the corpus has grown to the top 15 games from ten genres, and about 40 reviews per game. (Counting the successes) there are over 4 thousand reviews now to model from. I am currently pre-processing the larger data, lemmatizing and figuring out appropriate additions to my stop words, and min_df value to use (the above example was at 3) to create the largest possible, yet clearly defined topics to use. I will then use the pairwise distance to relate the strongest related review to its title, allowing me to recommend specific games. Past that I hope to create an interactive app or dashboard to allow this recommendation system to be functionally used.