Tilak D

March 10th 2017

**Proposal**: Create word vectors from a Mahabharata dataset to extract semantic similarities.

**Domain Background:**

According to Wikipedia “**Natural language processing (NLP)** is a field of [computer science](https://en.wikipedia.org/wiki/Computer_science), [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence), and [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics) concerned with the interactions between [computers](https://en.wikipedia.org/wiki/Computer) and [human (natural) languages](https://en.wikipedia.org/wiki/Natural_language) and, in particular, concerned with programming computers to fruitfully process large [natural language corpora](https://en.wikipedia.org/wiki/Corpus_linguistics).

Challenges in natural language processing frequently involve [natural language understanding](https://en.wikipedia.org/wiki/Natural_language_understanding), [natural language generation](https://en.wikipedia.org/wiki/Natural_language_generation) (frequently from [formal, machine-readable logical forms](https://en.wikipedia.org/wiki/Formal_language)), [connecting language and machine perception](https://en.wikipedia.org/wiki/Symbol_grounding_problem), [managing human-computer dialog systems](https://en.wikipedia.org/wiki/Dialog_system), or some combination thereof.”

The [***Mahabharata***](https://en.wikipedia.org/wiki/Mahabharata) is one of the two major [Sanskrit](https://en.wikipedia.org/wiki/Sanskrit_literature) [epics](https://en.wikipedia.org/wiki/Indian_epic_poetry) of [ancient India](https://en.wikipedia.org/wiki/History_of_India). The *Mahabharata* is an epic narrative of the [Kurukshetra War](https://en.wikipedia.org/wiki/Kurukshetra_War) and the fates of the [Kaurava](https://en.wikipedia.org/wiki/Kaurava) and the [Pandava](https://en.wikipedia.org/wiki/Pandava) princes. It also contains [philosophical](https://en.wikipedia.org/wiki/Hindu_philosophy) and devotional material, such as a discussion of the four "goals of life" or [*purusharthas*](https://en.wikipedia.org/wiki/Purusharthas). Among the principal works and stories in the *Mahabharata* are the [*Bhagavad Gita*](https://en.wikipedia.org/wiki/Bhagavad_Gita), the story of [Damayanti](https://en.wikipedia.org/wiki/Damayanti), an abbreviated version of the *Ramayana*, and the [Rishyasringa](https://en.wikipedia.org/wiki/Rishyasringa), often considered as works in their own right.

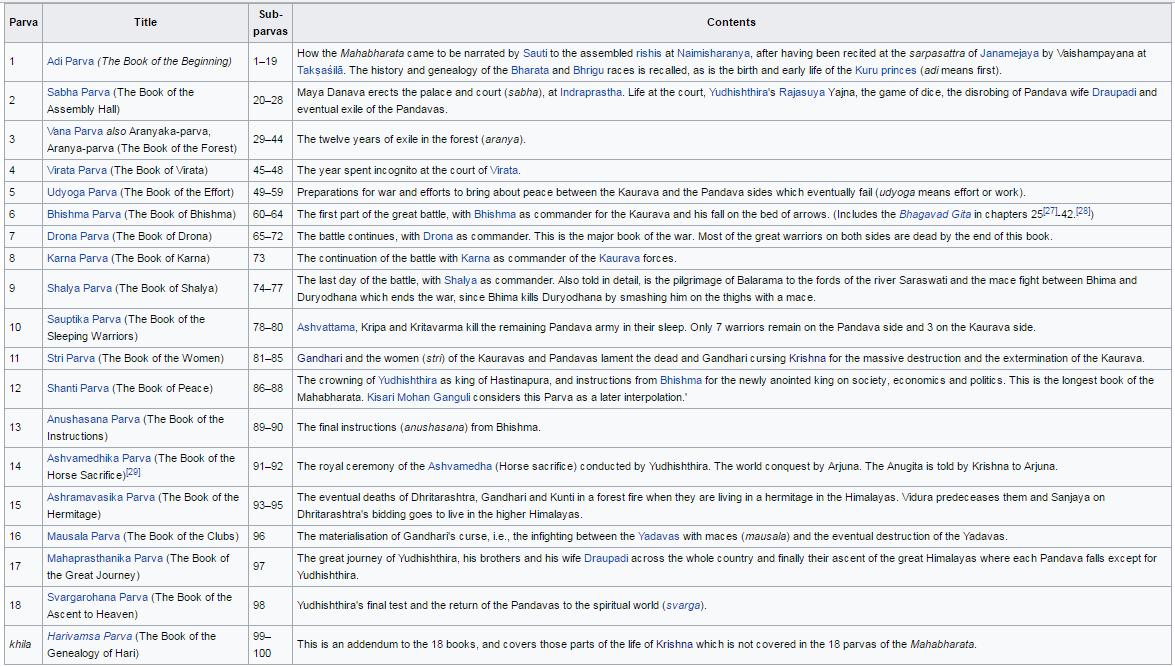
The *Mahabharata* is the longest known epic poem and has been described as "the longest poem ever written" Its longest version consists of over 100,000 [*shloka*](https://en.wikipedia.org/wiki/Shloka) or over 200,000 individual verse lines (each shloka is a couplet), and long prose passages. About 1.8 million words in total, the *Mahabharata* is roughly ten times the length of the [*Iliad*](https://en.wikipedia.org/wiki/Iliad) and the [*Odyssey*](https://en.wikipedia.org/wiki/Odyssey) combined, or about four times the length of the *Ramayana*, which makes it a huge dataset for using NLP.

**Problem Statement:**

In ancient times this knowledge used to pass along generations, but in this fast moving world, everyone needs answers easily and to be in their fingertips. Most of the relationships between characters is hard to remember, here NLP’s Semantic similarities come into play.

**Datasets and Inputs:**

Dataset is a set 18 text file, where in each text file is Parva (Which means book in Sanskrit). Below is an image taken from Wikipedia, having information of all 18 books.



Dataset was obtained from an online library, [Nitaaiveda](http://nitaaiveda.com/All_Scriptures_By_Acharyas/Historical_Works/Mahabharata.htm). Below is the statistics of the books, altogether combined.



This corpus of data will be used as an input to create word vectors using word2vec, with the help of NLTK to analyze semantic similarities.

**Solution Statement:**

As described above the corpus of data will be used as an input to create word vectors using word2vec, with the help of NLTK to analyze semantic similarities. The end solution of this project will be to analyze relationships and logics in the dataset. For example, Arjuna was the son of *Indra*- the king of celestials and Krishna was son of Vasudeva. If an input is given as Arjuna, Indra and Krishan, system should be capable to provide an answer as Vasudeva, based on the knowledge learnt using NLP.

**Benchmark Model:**

The problem which is being solved can only benchmarked based on the real info based on the book. As described in an example in Solution domain, Arjuna was the son of *Indra*- the king of celestials and Krishna was son of Vasudeva. If an input is given as Arjuna, Indra and Krishan, system should be capable to provide an answer as Vasudeva, based on the knowledge learnt using NLP. This result can only be objectively compared with the real facts.

**Evaluation Metrics:**

As explained in the previous section, the result obtained can only be objectively compared with the real facts. An evaluation metrics can only be a percentage of correct semantic obtained, which will be obtained through a sizable number of inputs.

**Project Design:**

Skeleton of the approach will be,

1. Create a dataset by converting corpus into sentences in turn into a bag of words.
2. Improve the dataset by removing the words and symbols that does not have meanings.
3. Build model by training word2vec and build a vocabulary.
4. The trained word vectors will be in a high dimension, example more than 200 dimension.
5. Using t-distributed stochastic neighbor embedding or t-SNE or PCA to reduce this higher dimension to a feasible, analyzable dimension size.
6. Train the dimensionality reduction algorithms to create a lower dimension dataset. Plot and analyze it for semantics.
7. For further analysis and to answer the problem statement, I am planning to use cosine similarity to answer similarity questions based on the dataset.