Machine Learning Engineer Nanodegree

Capstone Proposal

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, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages and, in particular, concerned with programming computers to fruitfully process large natural language corpora.

Challenges in natural language processing frequently involve natural language understanding, natural language generation (frequently from formal, machine-readable logical forms), connecting language and machine perception, managing human-computer dialog systems, or some combination thereof."

The ***Mahabharata*** is one of the two major Sanskrit epics of ancient India. The *Mahabharata* is an epic narrative of the Kurukshetra War and the fates of the Kaurava and the Pandava princes. It also contains philosophical and devotional material, such as a discussion of the four "goals of life" or *purusharthas*. Among the principal works and stories in the *Mahabharata* are the *Bhagavad Gita*, the story of Damayanti, an abbreviated version of the *Ramayana*, and the 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* 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* and the *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.

Parva	Title	Sub-parvas	Contents
1	Adi Parva *(The Book of the Beginning)*	1–19	How the *Mahabharata* came to be narrated by Sauti to the assembled rishis at Naimisharanya, after having been recited at the *sarpasattra* of Janamejaya by Vaishampayana at Takṣaśilā. The history and genealogy of the Bharata and Bhrigu races is recalled, as is the birth and early life of the Kuru princes (*adi* means first).
2	Sabha Parva (The Book of the Assembly Hall)	20–28	Maya Danava erects the palace and court (*sabha*), at Indraprastha. Life at the court, Yudhishthira's Rajasuya Yajna, the game of dice, the disrobing of Pandava wife Draupadi and eventual exile of the Pandavas.
3	Vana Parva *also* Aranyaka-parva, Aranya-parva (The Book of the Forest)	29–44	The twelve years of exile in the forest (*aranya*).
4	Virata Parva (The Book of Virata)	45–48	The year spent incognito at the court of Virata.
5	Udyoga Parva (The Book of the Effort)	49–59	Preparations for war and efforts to bring about peace between the Kaurava and the Pandava sides which eventually fail (*udyoga* means effort or work).
6	Bhishma Parva (The Book of Bhishma)	60–64	The first part of the great battle, with Bhishma as commander for the Kaurava and his fall on the bed of arrows. (Includes the *Bhagavad Gita* in chapters 25[27]-42.[28])
7	Drona Parva (The Book of Drona)	65–72	The battle continues, with Drona as commander. This is the major book of the war. Most of the great warriors on both sides are dead by the end of this book.
8	Karna Parva (The Book of Karna)	73	The continuation of the battle with Karna as commander of the Kaurava forces.
9	Shalya Parva (The Book of Shalya)	74–77	The last day of the battle, with Shalya as commander. Also told in detail, is the pilgrimage of Balarama to the fords of the river Saraswati and the mace fight between Bhima and Duryodhana which ends the war, since Bhima kills Duryodhana by smashing him on the thighs with a mace.
10	Sauptika Parva (The Book of the Sleeping Warriors)	78–80	Ashvattama, Kripa and Kritavarma kill the remaining Pandava army in their sleep. Only 7 warriors remain on the Pandava side and 3 on the Kaurava side.
11	Stri Parva (The Book of the Women)	81–85	Gandhari and the women (*stri*) of the Kauravas and Pandavas lament the dead and Gandhari cursing Krishna for the massive destruction and the extermination of the Kaurava.
12	Shanti Parva (The Book of Peace)	86–88	The crowning of Yudhishthira as king of Hastinapura, and instructions from Bhishma for the newly anointed king on society, economics and politics. This is the longest book of the Mahabharata. Kisari Mohan Ganguli considers this Parva as a later interpolation.'
13	Anushasana Parva (The Book of the Instructions)	89–90	The final instructions (*anushasana*) from Bhishma.
14	Ashvamedhika Parva (The Book of the Horse Sacrifice)[29]	91–92	The royal ceremony of the Ashvamedha (Horse sacrifice) conducted by Yudhishthira. The world conquest by Arjuna. The Anugita is told by Krishna to Arjuna.
15	Ashramavasika Parva (The Book of the Hermitage)	93–95	The eventual deaths of Dhritarashtra, Gandhari and Kunti in a forest fire when they are living in a hermitage in the Himalayas. Vidura predeceases them and Sanjaya on Dhritarashtra's bidding goes to live in the higher Himalayas.
16	Mausala Parva (The Book of the Clubs)	96	The materialisation of Gandhari's curse, i.e., the infighting between the Yadavas with maces (*mausala*) and the eventual destruction of the Yadavas.
17	Mahaprasthanika Parva (The Book of the Great Journey)	97	The great journey of Yudhishthira, his brothers and his wife Draupadi across the whole country and finally their ascent of the great Himalayas where each Pandava falls except for Yudhishthira.
18	Svargarohana Parva (The Book of the Ascent to Heaven)	98	Yudhishthira's final test and the return of the Pandavas to the spiritual world (*svarga*).
khila	*Harivamsa Parva* (The Book of the Genealogy of Hari)	99–100	This is an addendum to the 18 books, and covers those parts of the life of Krishna which is not covered in the 18 parvas of the *Mahabharata*.

Dataset was obtained from an online library, [Nitaaiveda](). Below is the statistics of the books, altogether combined.

Statistics:

Pages	461
Words	291,175
Characters (no spaces)	1,410,985
Characters (with spaces)	1,700,430
Paragraphs	2,585
Lines	18,634

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.