

Machine Learning Engineer Nanodegree

Capstone Proposal

Tilak D

March 15th 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.

By utilizing NLP, we can organize and structure knowledge of the huge Mahabharata to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis and topic segmentation, which will be helpful for extracting quick, short and concise answers.

Word2vec ([Ref\[1\]](#)) is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes a large corpus of text as input and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space ([Ref\[2\]](#)). Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

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 in lengthy novels are hard to remember for general public, here NLP's Semantic similarities come into play.

This corpus of data, of about 29100 words will be fed as input to the model to create word vectors and with the help of word2vec, we would analyze semantic similarities between characters, i.e. word vector representation will be created with neural network and by measuring similarity between similar words through cosine similarity to find the similarity between words. This will be used to answer questions related to relationships between characters in Mahabharata. 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.

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	<i>Adi Parva</i> (The Book of the Beginning)	1–19	How the <i>Mahabharata</i> came to be narrated by <i>Sauti</i> to the assembled <i>rishis</i> at <i>Naimisharanya</i> , after having been recited at the <i>śarapasattra</i> of <i>Janamejaya</i> by <i>Vaishampayana</i> at <i>Takṣaśāṭā</i> . The history and genealogy of the <i>Bharata</i> and <i>Bhṛigu</i> races is recalled, as is the birth and early life of the <i>Kuru</i> princes (<i>adi</i> means first).
2	<i>Sabha Parva</i> (The Book of the Assembly Hall)	20–28	Maya Danava erects the palace and court (<i>sabha</i>), at <i>Indraprastha</i> . Life at the court, <i>Yudhishtira's</i> <i>Rajasuya</i> Yajna, the game of dice, the disrobing of Pandava wife <i>Draupadi</i> and eventual exile of the Pandavas.
3	<i>Vana Parva</i> also <i>Aranyaka-parva</i> , <i>Aranya-parva</i> (The Book of the Forest)	29–44	The twelve years of exile in the forest (<i>aranya</i>).
4	<i>Virata Parva</i> (The Book of Virata)	45–48	The year spent incognito at the court of <i>Virata</i> .
5	<i>Udyoga Parva</i> (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 (<i>udyoga</i> means effort or work).
6	<i>Bhisma Parva</i> (The Book of Bhisma)	60–64	The first part of the great battle, with <i>Bhisma</i> as commander for the Kaurava and his fall on the bed of arrows. (Includes the <i>Bhagavad Gita</i> in chapters 25 ^[27] 42 ^[28])
7	<i>Drona Parva</i> (The Book of Drona)	65–72	The battle continues, with <i>Drona</i> 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	<i>Karna Parva</i> (The Book of Karna)	73	The continuation of the battle with <i>Karna</i> as commander of the <i>Kaurava</i> forces.
9	<i>Shalya Parva</i> (The Book of Shalya)	74–77	The last day of the battle, with <i>Shalya</i> 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	<i>Saṃvatsara Parva</i> (The Book of the Sleeping Warriors)	78–80	<i>Ashvattama</i> , <i>Kripa</i> and <i>Kritavarma</i> kill the remaining Pandava army in their sleep. Only 7 warriors remain on the Pandava side and 3 on the Kaurava side.
11	<i>Stri Parva</i> (The Book of the Women)	81–85	<i>Gandhari</i> and the women (<i>stri</i>) of the Kauravas and Pandavas lament the dead and <i>Gandhari</i> cursing Krishna for the massive destruction and the extermination of the Kaurava.
12	<i>Shanti Parva</i> (The Book of Peace)	86–88	The crowning of <i>Yudhishtira</i> as king of Hastinapura, and instructions from <i>Bhisma</i> for the newly anointed king on society, economics and politics. This is the longest book of the Mahabharata. <i>Kisari Mohan Ganguli</i> considers this Parva as a later interpolation.'
13	<i>Anushasana Parva</i> (The Book of the Instructions)	89–90	The final instructions (<i>anushasana</i>) from <i>Bhisma</i> .
14	<i>Ashvamedhika Parva</i> (The Book of the Horse Sacrifice) ^[29]	91–92	The royal ceremony of the <i>Ashvamedha</i> (Horse sacrifice) conducted by <i>Yudhishtira</i> . The world conquest by Arjuna. The <i>Anugita</i> is told by Krishna to Arjuna.
15	<i>Ashramavasika Parva</i> (The Book of the Hermitage)	93–95	The eventual deaths of <i>Dhritarashtra</i> , <i>Gandhari</i> and <i>Kunti</i> in a forest fire when they are living in a hermitage in the Himalayas. <i>Vidura</i> predeceases them and <i>Sanjaya</i> on <i>Dhritarashtra's</i> bidding goes to live in the higher Himalayas.
16	<i>Mausala Parva</i> (The Book of the Clubs)	96	The materialisation of <i>Gandhari's</i> curse, i.e., the infighting between the <i>Yadavas</i> with maces (<i>mausala</i>) and the eventual destruction of the <i>Yadavas</i> .
17	<i>Mahaprasthanika Parva</i> (The Book of the Great Journey)	97	The great journey of <i>Yudhishtira</i> , his brothers and his wife <i>Draupadi</i> across the whole country and finally their ascent of the great Himalayas where each Pandava falls except for <i>Yudhishtira</i> .
18	<i>Svargarohana Parva</i> (The Book of the Ascent to Heaven)	98	<i>Yudhishtira's</i> final test and the return of the Pandavas to the spiritual world (<i>svarga</i>).
<i>khila</i>	<i>Harivamsa Parva</i> (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 <i>Mahabharata</i> .

Dataset was obtained from an online library, [Nitaaveda](#). 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, of about 29100 words will be fed as input to the model to create word vectors using word2vec ([Ref\[3\]](#)), and with the help of NLTK, we would analyze semantic similarities between characters.

Solution Statement:

As described above the corpus of words will be used as an input to create word vectors ([Ref\[4\]](#)) using word2vec, with the help of t-SNE ([Ref\[5\]](#)) or PCA ([Ref\[6\]](#)) reduce the dimensions of the word vectors and finally use cosine similarity to analyze semantic similarities, i.e. to answer relationship questions based on the learning. The end solution of this project will be to analyze relationships and logics in the dataset.

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.

The real facts about the data set already exists, to benchmark the model I have compiled 23 relationship facts and will be adding few more as I build the model.

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. The model will be tested with an input containing a list of all the semantics obtained from Mahabharata in the below format.

{A} is related to {B}, as {C} is related to {D}

Here A, B and D would be the inputs and C will be the output provided by the model, based on the learning. Below is a subset of the compiled test semantics (Red words are the expected outputs),

Dhritarastra is related to Pandu, as Sahadeva is related to Nakula
Bhima is related to Arjuna, as Ambalika is related to Ambika
Pandur is related to Kunti, as Dhritarashtra is related to Gandhari
Bhima is related to Draupadi, as Arjuna is related to Chitrangada
Karna is related to Kunti, as Duryodhana is related to Gandhari
Bhima is related to Draupadi, as Arjuna is related to Subhadra
Yudhishthira is related to Kunti, as Duryodhana is related to Gandhari
Bhima is related to Kunti, as Nakula is related to Madri
Bhima is related to Draupadi, as Arjuna is related to Ulupi
Vichitravirya is related to Ambalika, as Vichitravirya is related to Ambika
Bhima is related to Ghatotkacha, as Arjuna is related to Abhimanyu
Bhima is related to Draupadi, as Arjuna is related to Draupadi

The output provided by the model is compared to the actual semantics and a percentage accuracy is calculated, i.e. (Total number of right answers given)/(Test size).

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. Train the above dimensionality reduction algorithms to create a lower dimension dataset. Plot and analyze it for semantics.
- 6) For further analysis and to answer the problem statement, I am planning to use cosine similarity, to assess similarities between 2 word vectors, to answer similarity questions on the 3rd word vector.

Extras:

Future improvements that I am planning to do:

- 1) Automatic summarization
- 2) Named entity recognition
- 3) Sentiment analysis
- 4) Topic segmentation