Clustering

May 12, 2018

1 Clustering

1.1 Clustering Task

- Use Affinity Propagraion from Clustering.jl, to cluster word2vec word embeddings, according to meaning.
- Done right this will seperate locations from sports
- Done finely and it will seperate ball-sports from other sports, and will seperate locations according to regions, etc
- Affinity propagraion requires a similarity matrix, which you can set as a negated distance matrix.
- For this you'll also want Distances.jl for all your distance metric needs.
- It is traditional with word2vec to use cosine distance.
- You will as also need to set each item's availability. This is the diagonal of the similarity matrix. Decreasing it roughly corresponds to decreasing the amount each node wants to be in a cluster on its own.

2 First we loadup some data

For the the example presented here, we will use a subhset of Word Embedding, trained using Word2Vec.jl. These are 100 dimentional vectors, which encode syntactic and semantic information about words.

You can download the datased from here, and load it up with JLD as shown below. (or just load it directly if you have cloned the notebooks)

```
In [ ]: using JLD
        embeddings = load("../assets/ClusteringAndDimentionalityReduction.jld", "em
```