Creating Semantic Spaces Using Docume

Yahya Emara, Tristan Weger, Rya

Task

Pick dataset to use
Overview Entire code
Create token parsing code
Create token filtering network
Embed tokens into high dimension semantic space
Create code and network to cluster tokens
Create a final tuned filter
Create a cognitive map for 2D data
Create a visual representation of ~50D data
Prepare Final Presentation
Determine what kind of machine learning method to use (supervised, reinforcement, unsupervised, transfer, etc.)
Run the clustering algorithm using the different data subset partitions.
Select a cross validation method to get a more accurate rating of the performance of each model
Interpret the clustering results and make adjustments if needed.
Determine optimal amount of vector space reduction that can be achieved without losing anything above minimal cl
Decide methodology of dimension reduction and document rationale behind choice.
Actual implementation of dimension reduction.
Compare application results when reducing to different numbers of dimensions.
Document if the optimal number of dimensions is relatively consistent across different document embedding techni
Design quality and informative visualizations for reduced 2D vector space.
Potential. Employ different method/tool in order to reduce dimensions. Convey whether results remain consistent o
Break document into word pairs
Break document into sentences
Filter data to keep relevant word tokens
Embed tokens into high dimensionality semantic vectors
Compare embedding methods and optimize using the best method

ent Clustering Effort Matrix

n Rubadue

Ryan % Work	Tristan % Work	Yahya % Work	Rough Total Task Hours
50	25	25	3
33	**33**	**33**	10
25	**50**	25	20
25	**50**	25	8
25	25	**50**	10
25	25	**50**	10
50	25	25	10
25	25	**50**	10
25	25	**50**	3
33	**33**	**33**	5
25	**50**	25	3
25	**50**	25	3
25	**50**	25	2
25	**50**	25	3
50	25	25	2.5
50	25	25	3
50	25	25	7
50	25	25	4
50	25	25	2
50	25	25	2.5
50	25	25	5
25	25	**50**	5
25	25	**50**	5
25	25	**50**	5
25	25	**50**	10
25	25	**50**	5