Data Warehousing and Data Mining (Major Project)

Deadline: November 17, 2016 11:59 PM

Instructions:

1. Each topic can be assigned to 3 teams at max on **first come first serve basis,** so fill the form with your preferences by **October 25, 11:59 PM**

2. This project will be evaluated based on Concept clarity (25%), Explanation and Results (25%), Paper presentation and Viva (50%).

NOTE1 : After the projects are allotted get in contact with respective TA's for further explanation.
NOTE 2: 5,6,7 and 8 project are along are the lines of the mini project (kaggle). These projects can be done using deep learning techniques. It is preferable, if you have laptop with Nvidia GPU of at least 1GB.

Only for Project 1 and 2

3. Each project has 8-9 questions.
4. No implementation is required. You just need to answer the questions given.
5. Each answer should be between 0.5 to 1.5 pages in length.
6. The final submission should be in standard IEEE 2-column format.
7. Wherever required use figures for illustrating the concept.
8. Figures should be typically 3inches X 3inches.
9. The final draft can be a maximum of 8 pages (excluding references) and a minimum of 6 pages (excluding references).
10. For each question we have clearly listed the references based on which you can answer.
11. Cite specific references in your paper apart from the ones provided if required.
12. Teams copying content directly from different available resources will be given straight **ZERO**. So it would be better that you understand the concept and then explain it in your own words.

INDEX

Project Number	Project Name	TA
1	Word2Vec and Word Embeddings	Saket
2	Neural Network Based Clustering	Saket
3	Quality Assessment of Wikipedia Articles	Saket
4	Comparing Anomaly detection Algorithms for Keystroke Dynamics	Saket
5	Identification and classification of keyphrases	Raghavendra
6	Extraction of relationship between two identified keyphrases	Raghavendra
7	Bond Liquidity Prediction	Raghavendra
8	Real-Time Crime Forecasting Challenge	Raghavendra
9	Method mention extraction from scientific research paper	Shikha
10	A data mining approach to analysis and prediction of movie ratings	Shikha
11	Twitter Sentiment Analysis	Shikha
12	Community Detection for testing hypothesis of dispersion similarity	Shikha
13	Community Detection in Social Networks	Madan

14	Understanding Latent Dirichlet Allocation (LDA)	Madan
15	Learning node embeddings in Networks using Deep Learning	Madan
16	Recommendation Systems	Madan

1. Word2Vec and Word Embeddings

Questions:

1. What are Word embeddings? What are its applications? [1], [2]
2. What is Word2vec? [3]
3. Explain the CBOW model used in Word2vec. [3], [6]
4. Explain the Skip-gram model used in Word2vec. [4]
5. Are the results of Word2vec training sensitive to parametrization? [2], [3], [5]
6. Can biases exist in Word Embeddings? How to tackle this problem? [7]
7. Can you extend the Skip-gram algorithm in [3] to make it learn representations for nodes in an arbitrary connected graph? Explain. [8]
8. How can you relate the Word2vec learning algorithm to matrix factorisation based methods? Are they both effectively learning the same representations? [9]

References:

1. https://en.wikipedia.org/wiki/Word_embedding
2. https://web.stanford.edu/~jurafsky/slp3/19.pdf
3. https://arxiv.org/pdf/1301.3781v3.pdf
4. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
5. https://arxiv.org/pdf/1402.3722v1.pdf
6. https://arxiv.org/pdf/1411.2738.pdf
7. https://arxiv.org/pdf/1607.06520.pdf
8. http://www.perozzi.net/projects/deepwalk/
9. https://levyomer.wordpress.com/2014/09/10/neural-word-embeddings-as-implicit-matrix-factorization/

2. Neural Network Based Clustering

Questions:

1. What is data clustering and why is it used? [1]
2. What are its applications? [2]
3. How lateral inhibition is related to clustering algorithms? [4]
4. Why do most clustering methods resort to local partitioning of data instead of global optimal partitioning? [5]
5. What are the initial conditions in clustering based on neural network? [6]
6. How the weight assigned to the connections between neurons? [7]
7. Write pseudo-code of neural network based clustering algorithm? [8]
8. When will the transmitting process in data clustering stops (terminating condition)? [9]
9. How to suppress noisy data from making its own classes? [10]

References:

[1] http://link.springer.com/chapter/10.1007/11840930_45#page-1
[2]http://popelka.ms.mff.cuni.cz/cerno/structure_and_recognition/files/kukacka_clustering_referat.pdf
[3] http://link.springer.com/referenceworkentry/10.1007/978-0-387-79948-3_1379
[4] http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5069742
[5]https://books.google.co.in/books?id=J5ZuCQAAQBAJ&dq=global+minimum+of+multiextremal+functional+clustering
[6] http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5069742
[7] http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5069742
[8] http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5069742
[9] http://link.springer.com/chapter/10.1007/11840930_45#page-1
[10] http://link.springer.com/chapter/10.1007/11840930_45#page-1

3. Quality Assessment of Wikipedia Articles

As Wikipedia became the largest human knowledge repository, quality measurement of its articles received a lot of attention during the last decade. Most research efforts focused on classification of Wikipedia articles quality by using a different feature set. However, so far, no golden feature set was proposed. In this task, you have to implement the research paper given in reference and try to reproduce only the non deep learning results (kNN etc). Bonus marks if you can reproduce results for Deep Learning techniques.

Reference: http://dl.acm.org/citation.cfm?id=2910917

4. Comparing Anomaly detection Algorithms for Keystroke Dynamics

Keystroke dynamics - the analysis of typing rhythms to discriminate among users—has been proposed for detecting impostors (i.e., both insiders and external attackers). Since many anomaly-detection algorithms have been proposed for this task, it is natural to ask which are the top performers (e.g., to identify promising research directions). In this task, you have to implement the research paper given in reference and try to reproduce the results.

Reference: http://www.cs.cmu.edu/~keystroke/KillourhyMaxion09.pdf

Dataset and Evaluation Script: http://www.cs.cmu.edu/~keystroke/

5. Identification and classification of keyphrases

Reference: https://scienceie.github.io/example.html (Subtask A and B)

6. Extraction of relationship between two identified keyphrases

Reference: https://scienceie.github.io/example.html (Subtask C)

7. Bond Liquidity Prediction

Reference:
https://www.hackerrank.com/contests/gs-quantify-2016/challenges/gs-quantify-2016-challenge-1

8. Real-Time Crime Forecasting Challenge

Reference:
http://nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx?utm_source=KDNuggets&amp;utm_medium=Ad&amp;utm_content=Email&amp;utm_campaign=ForecastingChallenge#importantdates

9. Method mention extraction from scientific research paper

Extraction of method phrases that contain an explicit mention of method keyword. Linguistic techniques as well as Statistical methods are expected to be used.

Reference: http://www.aclweb.org/anthology/C12-1074

10. A data mining approach to analysis and prediction of movie ratings

Perform relevance analysis to see what factors contribute most to a high rated movie, clustering to attempt to detect any relationships between the year a film is produced and its rating, and finally classification to attempt to classify the general rating of upcoming films based upon known information.

Reference:http://usir.salford.ac.uk/18838/1/Wessex_movie.pdf

11. Twitter Sentiment Analysis

Classification of tweets based on sentiment. The project should aim to use existing lexical resources as well as features that capture information about the informal and creative language used in microblogging.
Expectation: A hybrid approach using both corpus based and dictionary based methods to determine the semantic orientation.

Reference: http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf

12. Community Detection for testing hypothesis of dispersion similarity

Understand dispersion based similarity in a network. Implement and check if a detecting communities in such networks fits the hypothesis of dispersion based similarity.

Reference:
https://pdfs.semanticscholar.org/ea6a/254b119ea0351290509591dfbf94a764bf1e.pdf

13. Learning to Discover Circles in Social Networks (Community Detection)

Understand the concept of community detection in social networks. Implement and verify the results mentioned in [1], ONLY FOR FACEBOOK DATASET. Detailed explanation and further work on the paper can be found at [2].
Link to dataset - [3]. Python is the preferred language.

References:
[1] http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2012_0272.pdf
[2] https://cs.stanford.edu/people/jure/pubs/circles-tkdd14.pdf
[3] https://snap.stanford.edu/data/egonets-Facebook.html

14. Understanding Latent Dirichlet Allocation (LDA)

LDA is the most popular unsupervised technique to understand topic distributions in text documents. As part of the project, you need to understand the concept, mathematics and the working of the LDA, as explained in [1]. More details can be found at [2]. Video tutorials - [3].
No need to implement anything. A thorough viva would be conducted during the evaluation.
P.S. Paper [1] has 15978 citations till date, and that’s A LOT.

[1] http://ai.stanford.edu/~ang/papers/nips01-lda.pdf
[2] https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf
[3] http://videolectures.net/mlss09uk_blei_tm/

15. Learning node embeddings in Networks using Deep Learning

Deep Learning has done wonders in learning word embeddings. In recent years, it's applications for node embeddings has also been studied. In this project, you need to understand thoroughly the following two papers [1], [2]. Evaluation would be based on a thorough viva.

Interested people can go ahead and try implementing the paper [1] for the blogcatalog dataset [3]. Suggested languages would be Lua (Torch), Python (Tensorflow). Bonus marks for good attempts at implementing the paper.

References:

[1] http://www.www2015.it/documents/proceedings/proceedings/p1067.pdf
[2] https://arxiv.org/pdf/1403.6652.pdf
[3] http://socialcomputing.asu.edu/datasets/BlogCatalog3

16. Recommendation Systems

For this project, you are expected to understand the concepts of recommendation system. You can start from [1], but it is highly advised to explore.

Evaluation would be based on a thorough viva.

References:

[1] http://infolab.stanford.edu/~ullman/mmds/ch9.pdf