Register No: 17BIT0368
Name: Purohit Om hemantkumar
Category: III

8/08/2021

Review-1

Internship Project Abstract:

The project aims to aid Research Analysts, who go through various articles on the web and is tedious to keep a track on all of them, especially chrome tabs which use a lot of memory. Furthermore, not only will this project shrink the chrome tabs needed for the research, it contains topic clustering, NLP base summarization, and context of information retrieval

[redacted]

In doing so the whole project eases the workflow of the analysts at Twimbit and augments their productivity.

Literature Review

Author	Title	Technology/Techniques used	Abstract	Dataset
A.C. Santha Sheela ,Dr.C.Jayakumar	Comparative Study of Syntactic Search Engine and Semantic Search Engine: A Survey	Semantic and syntactic search	Search engine symbolizes an extremely powerful and valuable tool for fetching any sort of information from Internet. There has been numerous researches carried on search engines techniques, the major ones are syntactic and semantic. Referring to the Syntactic web, the results obtained are purely as per the keyword match. That is the query outputs numerous web pages against the keyword match that may not even be relevant or meaningful. Whereas, unlike the syntactic web, the semantic web is a revised or upgraded version of the web which produces quiet meaningful and specific output as it has the potential to comprehend the query effectively. Few examples of Semantic based search engines include Kosmix, Hakia, Cognition, Swoogle and Lexxe. Whereas syntactic based search engines are Google, Yahoo, Ask. The work performs a comparison amidst the performance of semantic and syntactic based search engine and evaluates them by employing certain queries.	Kosmix, Hakia, Cognition, Swoogle and Lexxe. Whereas syntactic based search engines are Google, Yahoo, Ask.
Jianyou Lv, Yuqian Wang	Semantic Information Detection of Webpage Based on Word Vector and Infomap	Regular expression, viterbi algorithm, word segmentation,ngram2vec,infomap clustering, multi-layer neural network	For Chinese web pages, we use regular expression and Viterbi algorithm to realize Chinese filtering and word segmentation, then use ngram2vec algorithm to get the word vector set of web page and pre train the word vector set of Baidu Encyclopedia. Baidu Encyclopedia word vector set is based on Infomap clustering algorithm to realize word vectorClustering and tagging types, training neural network through training data set and Baidu Encyclopedia corpus to determine the type of unknown web pages through neural network, and achieve the purpose of detecting the semantic information of unknown web pages. This algorithm is has few super parameters and high calculation efficiency. Experiments show that the accuracy of the trained neural network model reaches 96.73%, which can quickly and accurately identify the type of web page	Chinese webpages and Baidu encyclopedia

Laureta Hajderanj ,Isakh Weheliye, Daqing Chen	ANew Supervised t-SNEwith dissimilarity Measure for Effective Data Visualization and Classification	S-tSNE ,t-SNE,KNN,dimensionality reduction	In this paper, a new version of the Supervised t- Stochastic Neighbor Embedding (S-tSNE) algorithm is proposed which introduces the use of a dissimilarity measure related to class information. The proposed S-tSNE can be applied in any high dimensional dataset for visualization or as a feature extraction for classification problems. In this study, the S-tSNE is applied to three datasets MNIST, Chest x-ray, and SEER Breast Cancer. The two-dimensional data generated by the S-tSNE showed better visualization and an improvement in terms of classification accuracy in comparison to the original t- Stochastic Neighbor Embedding(t-SNE) method. The results from k-nearest neighbors (k-NN) classification model which used the lower dimension space generated by the new S-tSNE method showed more than 20% improvement on average in accuracy in all the three datasets compared with the t-SNE method. Iaddition, the classification accuracy using the S-tSNE for feature extraction was even higher than classification accuracy obtained from the original high dimensional d	MNIST, chest x-rays and SEER Breast cancer dataset
Changzhou Li, Yao Lu, Junfeng Wu, Yongrui Zhang, Zhongzhou Xia, Tianchen Wang, Dantian Yu, Xurui Chen, Peidong Liu, Junyu Guo	LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering	Latent Dirchlet Allocation(LDA), Word2Vec, Document clustering, PW-LDA	Clustering narrow-domain short texts, such as academic abstracts, is an extremely difficult clustering problem. Firstly, short texts lead to low frequency and sparseness of words, making clustering results highly unstable and inaccurate; Secondly, narrow domain leads to great overlapping of insignificant words and makes it hard to distinguish between sub-domains, or fine-grained clusters. The vocabulary size is also too small to construct a good word bag needed by traditional clustering algorithms like LDA to give a meaningful topic distribution. A novel clustering model, Partitioned Word2Vec-LDA (PW-LDA), is proposed in this paper to tackle the described problems. Since the purpose sentences of an abstract contain crucial information about the topic of the paper, we firstly implement a novel algorithm to extract them from the abstracts according to its structural features. Then high-frequency words are removed from those purpose sentences to get a purified-purpose corpus and LDA and Word2Vec models are trained. After combining the results of both models, we can cluster the abstracts more precisely. Our model uses abstract text instead of keywords to cluster because keywords may be ambiguous and cause unsatisfied clustering results shown by previous work. Experimental results show that the clustering results of PW-LDA are much more accurate and stable than state-of-the-art techniques	Academic abstracts, Wan Fang med Database
Jie Liu, Chun Yu ,Wenchang Xu, Yuanchun Shi	Clustering Web Pages to Facilitate Revisitation on Mobile Devices	VSM(vector space model), TF-IDF(term frequency inverse document frequency)	Due to small screens, inaccuracy of input and other limitations of mobile devices, revisitation of Web pages in mobile browsers takes more time than that in desktopbrowsers. In this paper, we propose a novel approach to facilitate revisitation. We designed AutoWeb, a system that clusters opened Web pages into different topics based on their contents. Users can quickly find a desired opened Web page by narrowing down the searching scope to a group of Web pages that share the same topic. Clustering accuracy is evaluated to be 92.4%and computing resource consumption was proved to be acceptable. A user study wasconducted to explore user experience and how much AutoWeb facilitates revisitation. Results showed that AutoWeb could save up a significant time for revisitation and participants rated the system highly.	Web pages, MIT articles, Financial pages, clothing brands

REFERENCES

[1]Sheela, A. S., & Jayakumar, C. (2019, March). Comparative Study of Syntactic Search Engine and Semantic Search Engine: A Survey. In 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (Vol. 1, pp. 1-4). IEEE.

[2]Wang, Y., & Lv, J. (2020, July). Semantic Information Detection of Webpage Based on Word Vector and Infomap. In 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 293-297). IEEE.

[3]Hajderanj, L., Weheliye, I., & Chen, D. (2019, April). A new supervised t-SNE with dissimilarity measure for effective data visualization and classification. In Proceedings of the 2019 8th International Conference on Software and Information Engineering

[4] Li, C., Lu, Y., Wu, J., Zhang, Y., Xia, Z., Wang, T., ... & Guo, J. (2018, April). LDA meets Word2Vec: a novel model for academic abstract clustering. In Companion proceedings of the the web conference 2018 (pp. 1699-1706).

[5] Liu, J., Yu, C., Xu, W., & Shi, Y. (2012, February). Clustering web pages to facilitate revisitation on mobile devices. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (pp. 249-252).

Detailed Design

Modules:

- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]

Database: Raw File Structure in AWS using CSVs

Hardware requirements: NONE

Software requirements:

- Chrome
- Flask
- Postman
- Dash/Plotly
- NLP Libraries(NlTK,BERT..etc)
- Selenium and Beautifulsoup

Mini Project with guide Abstract

The biggest problem we face nowadays is not knowing whether it is going to rain or not. Imagine planning for something big, and all your plan gets cancelled just because it starts raining. Sounds pretty irritating, doesn't it? Well, now we have a solution to this problem and we call it the Rain Predictor. Rain Predictor is a software which will help us predict whether it is going to rain on a particular day or not. Rain Prediction will be done on the basis of different factors which affect the rain conditions, such as temperature, evaporation, humidity, wind speed, etc. On the basis of the available dataset on Vellore, which is timeseries format and raw data requiring preprocessing. We are going to do the prediction of rain on using time series forecasting techniques such as **ARIMA, Exponentional Smoothening and other autoregressive models**

Literature Survey

Title	Dataset	Abstract	Parameters	Advantages of the model	Limitations of the model
Prediction of Rain Attenuation Statistics from Measured Rain Rate Statistics using Synthetic	Rain rate data are taken from ITU-R data bank for different tropical and temperate locations to show the applicability of	Prediction of signal attenuation due to rain are important in the conception of microwave and millimetre wave communication	Site location, Latitude, Longitude, Elevation, Frequency	The present study shows how SST can be successfully used to convert measured rain rate statistics	Normally a 2.2648 accuracy is a good accuracy but since this is managing realtime disasters we require

Rain rate and rain attenuation prediction with experimental rain attenuation efforts in south-western Nigeria	This data is available at the Kitami Institute of Technology databank	Rain induced attenuation is a prominent loss factor for communication system design in the terrestrial and satellite- earth links. Its severity is more pronounced at frequencies above 10GHz [1]	The prominent ITU - R rain rate model as detailed in is based on the use of meteorological parameters available from ITU's 3M Group website. The Kitami rain rate Distribution model employs two regional	The ITU and RH models show good performance at low rain rates while Kitami model shows the worst result for the location. Fig . 3 shows the predicted rain attenuatio n at 12.736 GHz, 12.522GHz,	The model works at Certain frequency Only such as 31.4 GHZ
Novel integration-time conversion of rain-rate statistics for rain attenuation prediction models	A disdrometer has been used for rain accumulation measurements Thirty rain events during 2011-2012 are considered``	Rain attenuation prediction model is important for both satellite and terrestrial communication s.	The instability parameters are estimated from radiometric data to point the development of atmospheric	The nowcasting technique is, therefore, able to predict both rain occurrence and rain accumulatio n.	We get results at 22.24, 23.8, 26.4 and 31GHz but only optimal and consiferable is at 31.4Ghz
A model of rain attenuation in Ka band based on the Wiener prediction	DAH is a prediction model, which is proposed by Allnutt, Dissanayake and Haidara after analyzing the data that come from series of experiments based on INTEL SAT satellite system,	In this paper a new rain attenuation model based on Wiener prediction is established after analyzing the DAH model in Ka band.	Because of more parameters More complex process of calculation and the need of renewing all the parameters	In this paper, a new rain attenuation model is introduced based on the analysis of DAH model. Simulatio n results show that this new model can achieve the same effect with DAH model	Besides, this new model of 3th order has only 3 parameters, and there are no close relationshi p between the parameter s and frequency.
A New Rain Attenuation Prediction Model for the Earth-Space Links	Based on the measurement data by Meteorological radar,a rain attenuation prediction model was	Earth-space communicationOn systems arenow utilizing the Ku-and Ka- frequency	Twelve parameters sets, one for each month of the year, are available. The model	A new rain attenuation prediction model for the earth- space links is	over various Ranges of latitudes, frequencies, and

REFERENCES

[1] Nandi, D. (2018, November). Prediction of Rain Attenuation Statistics from Measured Rain Rate Statistics using Synthetic Storm Technique for Micro and Millimeter wave Communication Systems. In 2018 IEEE MTT-S International Microwave and RF Conference (IMaRC) (pp. 1-4). IEEE.

[2]Ibiyemi, T. S., Ajewole, M. O., Ojo, J. S., & Obiyemi, O. O. (2012, November). Rain rate and rain attenuation prediction with experimental rain attenuation efforts in south-western Nigeria. In 2012 20th Telecommunications Forum (TELFOR) (pp. 327-329). IEEE

[3]Thiennviboon, P., Intarawichian, S., Zhao, Z. W., Lin, L. K., & Lu, C. S. Novel integration-time conversion of rain-rate statistics for rain attenuation prediction models. In 2017 International Symposium on Antennas and Propagation (ISAP) (pp. 1-2). IEEE.

[4]Xinyu, D., Yong, S., & Yanling, W. (2010, November). A model of rain attenuation in Ka band based on the Wiener prediction. In Proceedings of Papers 5th European Conference on Circuits and Systems for Communications (ECCSC'10) (pp. 264-267). IEEE.

[5]Lu, C. S., Zhao, Z. W., Wu, Z. S., Lin, L. K., Thiennviboon, P., Zhang, X., & Lv, Z. F. (2018). A new rain attenuation prediction model for the earth-space links. IEEE Transactions on Antennas and Propagation, 66(10), 5432-5442.

Detailed Design

Modules:

- Data Aggregator(Excel files)
- Preprocessing data
- EDA(Exploratory Data Analysis)
- Detect and remove Trend, Seasonality as per Time Series standards
- Final Prediction using different algorithms

Database: Excel Files

Hardware requirements: NONE

Software requirements:

- Pandas
- Numpy and core preprocessing libraries
- Libraries such as Statistical models, Random Forest, Stacked ensemble, linear regression models
- Dashboard (Dash/plotly or D3.js)