TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 6649 - 6652

Text Mining and Sentimental Analysis on Comments using Machine Learning and NLP

Dharmapuri Saivamshi[1], V. Karthick[2], S. Magesh[3]
[1]UG Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamilnadu, India- 602 105
[2]Assistant Professor, Department of Cloud Computing and Information Sciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamilnadu, India-602 105
[3]Professor, Department of Cloud Computing and Information Sciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamilnadu, India-602105
[1]dsaivamshi123@gmail.com , [2]karthickv.sse@saveetha.com, [3]magesh.sse@saveetha.com

Article Info
Volume 82
Page Number: 6649 - 6652
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 01 February 2020*

Abstract
Now-a-days video streaming platforms like YouTube, twitch on-line etc. but the downside comes with the quantity of users and big information that is been generated through comments, truth essence of technology engineering is finding real-life problems. Throughout this project, we have a tendency to tend to creating associate algorithmic rule that collects and analyses the comments in glorious internet websites like Facebook, YouTube and various platforms. We have a tendency to tend to perform a sentimental analysis and word frequency over the collected information. There square measure some ways that to urge the output like internet extensions or an online web site. That the account holders will have a firm and clear arrange that viewers ought to expresses with relevance the content that is been uploaded by the admin. With the help of Python methodology like Scraping, the data collected from the comments square measure analyzed and final results square measure provided. For this project YouTube is been elite as a result of the bottom platform for development and aggregation information required as a result of it holds the foremost vital shopper or shopper information at intervals the kind of channel admins and viewer.

Keywords*:* Scraping, Tendency, Comments.

1. Introduction

Natural Language method, or natural language processing for temporary, is usually made public as a result of the automated manipulation of language, like speech and text, by code. language refers to the approach we have a tendency to tend to, humans, communicate with each other. Namely, speech and text. we have a tendency to tend to square measure encircled by text. the last word objective of natural language processing is to browse, decipher, understand, of the human languages in AN passing manner that is valuable. Most natural language processing techniques trust machine learning to derive which implies from therefore machine learning has become one all told rock bottom law of this project.

I have selected YouTube as surroundings to work on as a results of YouTube is go together with video-sharing platform headquartered in San Bruno, California. three former PayPal employees—Chad Hurley, Steve Chen, and Jawed Karim—created the service in Gregorian calendar month 2005.

It creates loads of info that's applicable for the algorithmic rule.

January-February 2020
ISSN: 0193-4120 Page No. 6649 - 6652

Human languages

The main drawback that comes with language is language ambiguity.

Example

The goat is prepared to eat.
The higher than example will mean in 2 ways that
1. The goat is grilled.
2. Goat itself is prepared to eat food.

The major issue with this is often laptop or the program couldn't be able to perceive the situation if we have a tendency to directly pass the concerning sentence while not process it. so methodology of natural language processing kicks in to save lots of the day.

2. Methodology

Bag of Words

Bag of words is additionally referred to as as a model however, in my opinion its not a model however additional that of a technique of to interrupt the sentence into individual works that build a pile of words to perform sure operations.

Example

Sai vamshi could be a smart boy.
After applying bag of words to the higher than sentence than output can be:

- Sai
- Vamshi
- Is
- A
- Good
- Boy

Tokenization

Is the method of segmenting running text into sentences and words. In essence, it's the task of cutting a text into items referred to as tokens, and at identical time discard sure characters, like punctuation. Following our example, the results of tokenization would be:

Example

Saveetha college of engineering could be a one among the best, smart and schools of saveetha university and it's Communications Security Establishment, ECE and Mech etc as departments. however Communications Security Establishment is simply too good!

After Tokenization

The tokenization process are repeatedlysuggestivelytricky once speaking medical subject text domains that comprise several hyphens, parentheses, and alternate punctuation marks.

Figure 1: After Tokenization

Stop Words

Stop words square measure most typically occurring words in any language that doesn't contribute something for language analysis and process.

TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 6649 - 6652

Few stop words of NLTK: {'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there',
'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some',
'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's',
'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we',
'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this',
'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all',

Figure 2 : Stop Words

Stemming

Refers to the method of slicing the tip or the start of words with the intention of removing affixes (lexical additions to the foundation of the word).

Affixes that square measure hooked up at the start of the word square measure referred to as prefixes (e.g. "astro" within the word "astrobiology") and also the ones hooked up at the tip of the word square measure referred to as suffixes (e.g. "ful" within the word "helpful").

Like, liked, seemingly -> Like

Kindest, kinder -> kind.

Lemmatization

Has the goal of decreasing a phrase to its base type and grouping along totally different styles of equal word. For instance, verbs in past rectangular measure was gift (e.G. "went" is modified to "go") and synonyms square measure unified (e.G. "best" is modified to "good"), consequently standardizing phrases with comparable aiming to their root. even though it seems closely associated with the stemming method, lemmatization uses a special approach to succeed in the foundation styles of words.

Topic Modeling

Is as a way for uncovering hidden systems in sets of texts or documents. In essence it clusters texts to get latent subjects supported their contents, process person words and distribution them values supported their distribution. This technique is based at thethat every report includes a combination of subjects which each topic includes a collection of phrases, which implies that if we will spot those hidden subjects we will liberate the which means that of our texts.

Topic modeling is rather beneficial for classifying texts, building recommender systems (e.G. to suggest you books supported your beyond readings) or maybe police investigation trends in on-line publications.

Reason why ML? why not DL:

ML stands for machine learning and decilitre stands for Deep learning.

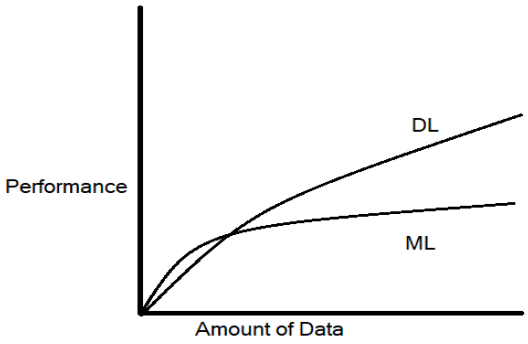

Figure 3: Deep Learning

According to graph on top of we are able to establish that performance of deciliter will increase with the rise in information. As a paradigm we tend to square measure aiming to work with significantly touch of information therefore milliliter may be a good match.

Future Works

Way forward for this project is attached building a deep learning network so the accuracy and performance increase's that lead a good advantage as a result of the comments size is dynamic ample comments square

January-February 2020
ISSN: 0193-4120 Page No. 6649 - 6652

measure been denote daily so this sort of situations square measure good or Deep learning.

Due to lack of information and time I even have hand-picked to create a millilitre model and building a deep learning Network may be a nice leap of development.

3. Conclusion

I herewith conclude this project will open a replacement manner of understanding immense quantity of information and cut back and even replace the human effort within the field of comments analysis. Comment analysis is being a pricey method however with the assistance of this project each YouTube will benefit of technology and improve their video content, develop their channel.

References

[1] Goldberg, Yoav (2016). A Primer on Neural Network Models for Natural Language Processing. Journal of Artificial Intelligence Research 57 (2016) 345–420

[2] Ian Goodfellow, Yoshua Bengio and Aaron Courville. Deep Learning. MIT Press.

[3] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu (2016). https://arxiv.org/abs/1602.02410 Exploring the Limits of Language Modeling

[4] Do Kook Choe and Eugene Charniak (EMNLP 2016). https://aclanthology.coli.uni-saarland.de/papers/D16-1257/d16-1257 Parsing as Language Modeling

[5] Vinyals, Oriol, et al. (NIPS2015). https://papers.nips.cc/paper/5635-grammar-as-a-foreign-language.pdf

[6] Asada, M.; Hosoda, K.; Kuniyoshi, Y.; Ishiguro, H.; Inui, T.; Yoshikawa, Y.; Ogino, M.; Yoshida, C. (2009). "Cognitive developmental robotics: a survey". IEEE Transactions on Autonomous Mental Development. 1 (1): 12–34. doi:10.1109/tamd.2009.2021702.

[7] "ACM Computing Classification System: Artificial intelligence". ACM. 1998. Archived from the original on 12 October 2007. Retrieved 30 August 2007.

[8] Goodman, Joanna (2016). Robots in Law: How Artificial Intelligence is Transforming Legal Services (1st ed.). Ark Group. ISBN 978-1-78358-264-8.

[9] Albus, J. S. (2002). "4D/RCS: A reference model architecture for intelligent unmanned ground vehicles". In Gerhart, G.; Gunderson, R.; Shoemaker, C. (eds.). 4-D/RCS: A Reference Model Architecture for Intelligent Unmanned Ground Vehicles (PDF). Proceedings of the SPIE AeroSense Session on Unmanned Ground Vehicle Technology. Unmanned Ground Vehicle Technology IV. 3693. pp. 11–20. Bibcode: 2002SPIE.4715. 303A. CiteSeerX 10.1.1.15.14. doi:10.1117/12.474462. Archived from the original (PDF) on 25 July 2004.

[10] Aleksander, Igor (1995). Artificial Neuroconsciousness: An Update. IWANN. Archived from the original on 2 March 1997. BibTex Archived 2 March 1997 at the Wayback Machine.

[11] Bach, Joscha (2008). "Seven Principles of Synthetic Intelligence". In Wang, Pei; Goertzel, Ben; Franklin, Stan (eds.). Artificial General Intelligence, 2008: Proceedings of the First AGI Conference. IOS Press. pp. 63–74. ISBN 978-1-58603-833-5.

[12] "Robots could demand legal rights". BBC News. 21 December 2006. Retrieved 3 February2011.

[13] Brooks, Rodney (1990). "Elephants Don't Play Chess" (PDF). Robotics and Autonomous Systems. 6 (1–2): 3–15. CiteSeerX 10.1.1.588.7539. doi:10.1016/S0921-8890(05)80025-9. Archived (PDF) from the original on 9 August 2007.

[14] Brooks, R. A. (1991). "How to build complete creatures rather than isolated cognitive simulators". In VanLehn, K. (ed.). Architectures for Intelligence. Hillsdale, NJ: Lawrence Erlbaum Associates. pp. 225–239. CiteSeerX 10.1.1.52.9510.

[15] Buchanan, Bruce G. (2005). "A (Very) Brief History of Artificial Intelligence" (PDF). AI Magazine: 53–60. Archived from the original (PDF) on 26 September 2007.

[16] Butler, Samuel (13 June 1863). "Darwin among the Machines". Letters to the Editor. The Press. Christchurch, New Zealand. Retrieved 16 October 2014 – via Victoria University of Wellington.