COMP5423 Internet Infrastructure and Protocols

Survey on Text Classification

Written Report

**Group Members & Contribution**

|  |  |
| --- | --- |
| Student ID | Name |
| 19074889G | WONG CHI HANG |

Table of Contents

[Task Definition 3](#_Toc70499856)

[Single Label Classification 3](#_Toc70499857)

[Multi-Label Classification 3](#_Toc70499858)

[Binary Classification 3](#_Toc70499859)

[Challenges 4](#_Toc70499860)

[Ambiguity 4](#_Toc70499861)

[Entity identification 4](#_Toc70499862)

[Lexical Gap 4](#_Toc70499863)

[Typical methods 4](#_Toc70499864)

[Tokenization 4](#_Toc70499865)

[Preprocessing 5](#_Toc70499866)

[Word Representation 5](#_Toc70499867)

[Classifier 6](#_Toc70499868)

[State-Of-The-Arts 6](#_Toc70499869)

[Transformer 6](#_Toc70499870)

[Bert 6](#_Toc70499871)

[XLNet 7](#_Toc70499872)

[ULMFit 7](#_Toc70499873)

[Application scenarios 7](#_Toc70499874)

[Email Filtering 7](#_Toc70499875)

[News Filtering 7](#_Toc70499876)

[Conclusion 8](#_Toc70499877)

[Reference 9](#_Toc70499878)

Task Definition

Text classification is a task to classify a text into different categories. The text can be a paragraph or a short sentence such as news title or email subject. [1] According to the target categories, text classification can be categorized as below types.

Single Label Classification

In this classification, a text is categorized to a single label. For example, a text can be categorized as “good”, “bad” or “average” etc. For a special situation that only two classes available for the text to be categorized, it is called binary classification. For example, categorize if an email is a spam or not by the email subject.

Multi-Label Classification

In this classification, a text is categorized to a multi-label. For example, a tweet may contain several tags.

Challenges

For the authentication, SFTP supports SSH Key authenticate, while FTPS supports X.509 certificate.

FTPS supports anonymous login, however, SFTP must be an account for login.

FTPS supports server-to-server copy but SFTP is not.

Ambiguity

Same as other NLP applications, lexical and semantic Ambiguity also are challenging issues to question answering system.

Entity identification

Entity identification is defined as a problem of identifying an object instance from different data sources that correspond to the same real-world entity.For example, from the MCTest dataset, story mc160.dev.1, question 1, "Who went to the beach with Sally?". The answer is "her brothers, mother, and father". The most related sentence in the story is "Sally went to the beach with her family in the summer as well." Actually, the family and Sally's brothers, mother, and father are referring to the same entities.

Lexical Gap

Lexical gap is refering to the problem that a word exists in a language but not exist in other language. For example, there is no word for the extra weight people gain from emotional binge eating in English. Due to the lexical gap, different ways may be used to express same meaning which increases the difficulty for the system to find out the correct answer for a question.

Typical methods

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

[4] – Related work

Fig [3]

Tokenization

Rule base Tokenizer

NLTK tokenizer

Statistical base Tokenizer

Spacy

All connection can be successfully created, expect the connection to FTPS on Linux server.

Preprocessing

Lower Case and Trim

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Stop Words Removal

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Stemming and Lemmatization

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Word Representation

Bag of Words

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Tf-idf

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Word Embedding

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

1. Word to Vector
2. GloVe
3. fastText

Classifier

To see the what are the difference in server-side implement, FTPS and SFTP server were implemented in Windows and Linux environment. The servers were set in Virtual Machine by using oracle virtual box.

State-Of-The-Arts

Transformer

In client-side implementation, Java is used as the programming language for the development.

Two java class are developed. One for FTPS, and the other for SFTP.

Bert

For the FTPS implementation, Apache Commons Net™ library is used to access ftps server.

Below is the execution result that upload and download completed successfully.

XLNet

For the FTPS implementation, Apache Commons Net™ library is used to access ftps server

ULMFit

In client-side implementation, Java is used as the programming language for the development.

Two java class are developed. One for FTPS, and the other for SFTP.

Application scenarios

From the study and implementation have done earlier, it seems that using SFTP in Windows environment is the most easy and efficient way for file transfer. To prove that, I have tried to implement a solution for a case scenario by using spring integration in Java.

Email Filtering

As the scheduler application was running, data files were generated at sftp download folder and ready for client to download.

News Filtering

As the scheduler application was running, data files were generated at sftp download folder and ready for client to download.

SftpDemo Module

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Also, uploaded to the sftp server.

Conclusion

Below files are submitted via blackboard.

1. FTPS\_and\_SFTP\_Comparison\_and\_Implementaion\_Group34\_19074889G.ppt
2. FTPS\_and\_SFTP\_Comparison\_and\_Implementaion\_Group34\_19074889G docx
3. FTPS\_and\_SFTP\_Comparison\_and\_Implementaion\_Group34\_19074889G\_workspace.zip

Reference

1. Techniques for text classification: Literature review and current trends

Microsoft. (2019). *New-SelfSignedCertificate.* [online] Available at: https://docs.microsoft.com/en-us/powershell/module/pkiclient/new-selfsignedcertificate?view=win10-ps [Accessed 13 May. 2020].

Computing for geeks. (2019). *Install and Configure OpenSSH Server on Windows Server 2019.* [online] Available at: https://computingforgeeks.com/install-and-configure-openssh-server-on-windows-server [Accessed 13 May. 2020].

Microsoft. (2019). *OpenSSH Server Configuration for Windows 10 1809 and Server 2019.* [online] Available at: https://docs.microsoft.com/en-us/windows-server/administration/openssh/openssh\_server\_configuration [Accessed 13 May. 2020].

server-world.info. (2019). *OpenSSH : SFTP only + Chroot.* [online] Available at: https://www.server-world.info/en/note?os=Ubuntu\_18.04&p=ssh&f=5 [Accessed 13 May. 2020].

How to forge. (2019). *How to install and configure VSFTPD.* [online] Available at: https://www.howtoforge.com/tutorial/how-to-install-and-configure-vsftpd/ [Accessed 13 May. 2020].