Company Risk Analyser in IBM Bluemix

Ting Yu ting.yu@sv.cmu.edu

Description

• Use IBM’s latest cloud platform – Bluemix, and its cloud services
• Use Restful API and MongoDB
• Automatically crawling the 10-K forms of public companies from Internet
• Get the “Risk Factors” for the companies by extracting the keywords for different types of risk
• Visualized the comparison of risk levels for different companies by years

Challenges

• Bluemix was in Beta version when the project started, the cloud services in Bluemix keep updating
• Not easy to make the web crawler effective, robust and stable
• Hard to define financial risk types and make a good keyword set for each type of risk
• Difficulty in the design of the algorithm for analyzing risk factors and getting risk levels

Solution – crawling, storage, parsing and visualization

Crawling → Storage → Parsing → Deploy & Visualization

• Crawling & storage: use crawler to download 10-k documents for all public companies and store the data into MongoDB
• Parsing: Define 25 types of financial risks and the keyword set for each type of risk. Extract keywords from the “Risk Factor” sections and find risks for the companies.
• Deploy & Visualization: Build a website using JavaEE and deploy it in Bluemix. Visualize comparison of the risks levels by years and for different companies

System Architecture

References

* https://ace.ng.bluemix.net/
* Highly extensible, highly scalable Web crawler
* http://nutch.apache.org/
* Natural Language Toolkit
* http://www.nltk.org/?

Advisor: Jia zhang
Sponsor: Nianjun Zhou

GH Wang gonghan.wang@sv.cmu.edu
Jacky Chen jacky.chen@sv.cmu.edu

jia.zhang@sv.cmu.edu
jzhou@us.ibm.com