Python package

0. pdfPlumber （从PDF中提取文字）

- 一个教程 （带有代码）

https://medium.com/analytics-vidhya/how-to-easily-extract-text-from-any-pdf-with-python-fc6efd1dedbe

1. FinBERT from HuggingFace transformers library

- 一个简介 + 案例 （带代码）：可以直接抄这个

https://wandb.ai/ivangoncharov/FinBERT_Sentiment_Analysis_Project/reports/Financial-Sentiment-Analysis-on-Stock-Market-Headlines-With-FinBERT-Hugging-Face--VmlldzoxMDQ4NjM0

- Hugging Face website : API， 包含了NLP需要的全部内容； 包括 tokenization; language model; token classification

https://huggingface.co/docs/transformers/index

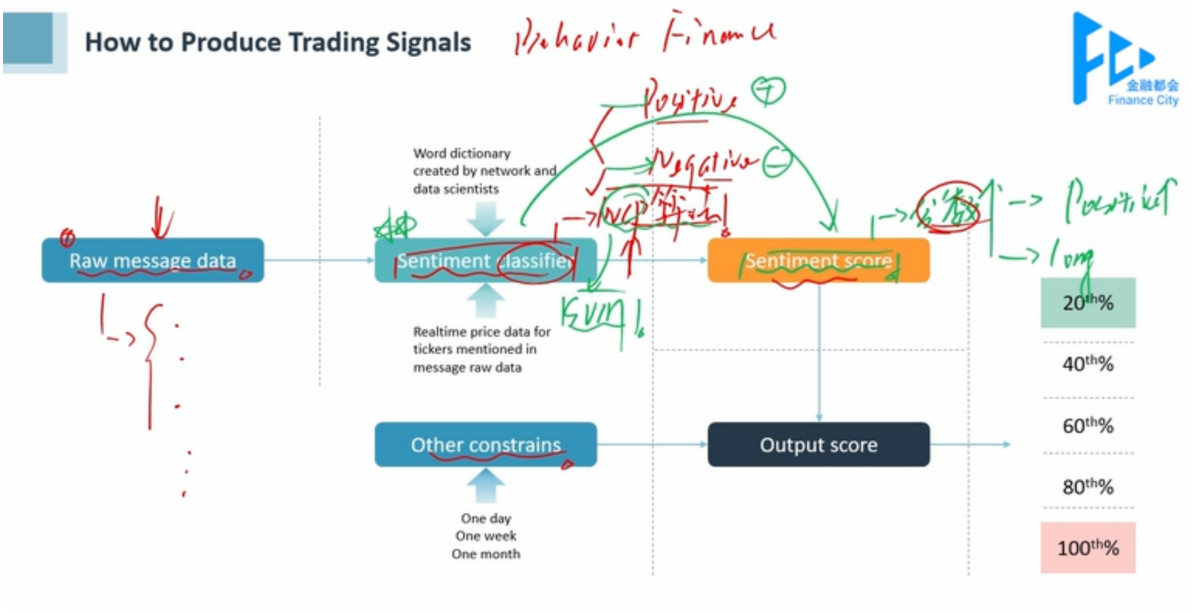
2. BERT (作为备选的语料库)

3. NLTK: 用于Raw_text ->word_list （作为备选）

- 一个简介：https://zhuanlan.zhihu.com/p/98808960
- 教程： 包括lexicon, processing raw text, categorizing and tagging words https://www.nltk.org/book/
- Doc: https://www.nltk.org/api/nltk.html
- 一个流程清晰的案例（带代码）： NLTK 做期权的舆情分析 https://towardsdatascience.com/a-step-by-step-tutorial-for-conducting-sentiment-analysis-a7190a444366

其它：

- 舆情分析 — 量化投资 流程



一个 sentiment analysis strategy

里面的图片会很有帮助

https://www.quantstart.com/articles/sentiment-analysis-trading-strategy-via-sentdex-data-in-qstrader/