

Pivotal.



Pivotal Greenplum: GPText

Pivotal's In-Database Text Analytics for Big Data

挖掘非结构化数据的金矿

*“...most industry experts agree that **80% to 90% of the world’s data is unstructured data**. Of these unthinkably vast stores, only 0.5% is effectively analyzed and used today.*

*In the business world, most unstructured data lies in **customer-related text**. However, most organizations don’t know how to efficiently extract predictive elements from unstructured customer data. But, done right, extracting valuable predictive insights from huge quantities of text takes just **seconds**.*

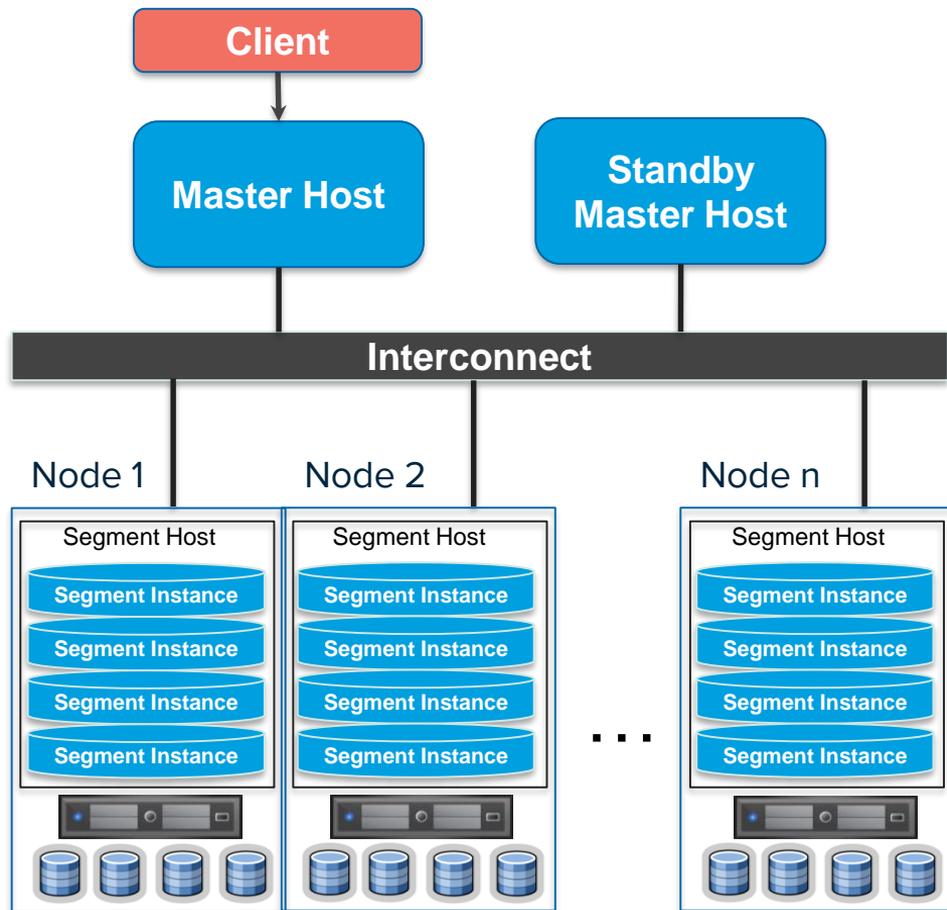
- Osvaldo Driollet (PhD), Sr. Data Scientist, FICO

A dark, atmospheric photograph of a city street, likely in New York City, featuring tall buildings and a prominent skyscraper in the center. The image is dimly lit, with a blueish-grey color palette. The text 'GPTText 架构简介' is overlaid in the center in a bright white font.

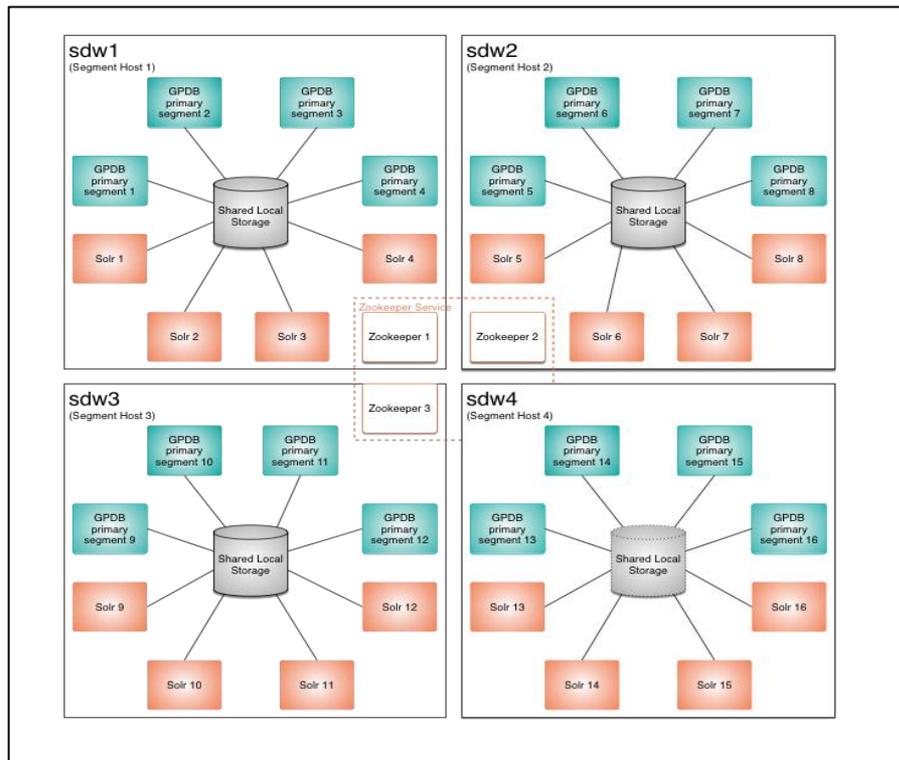
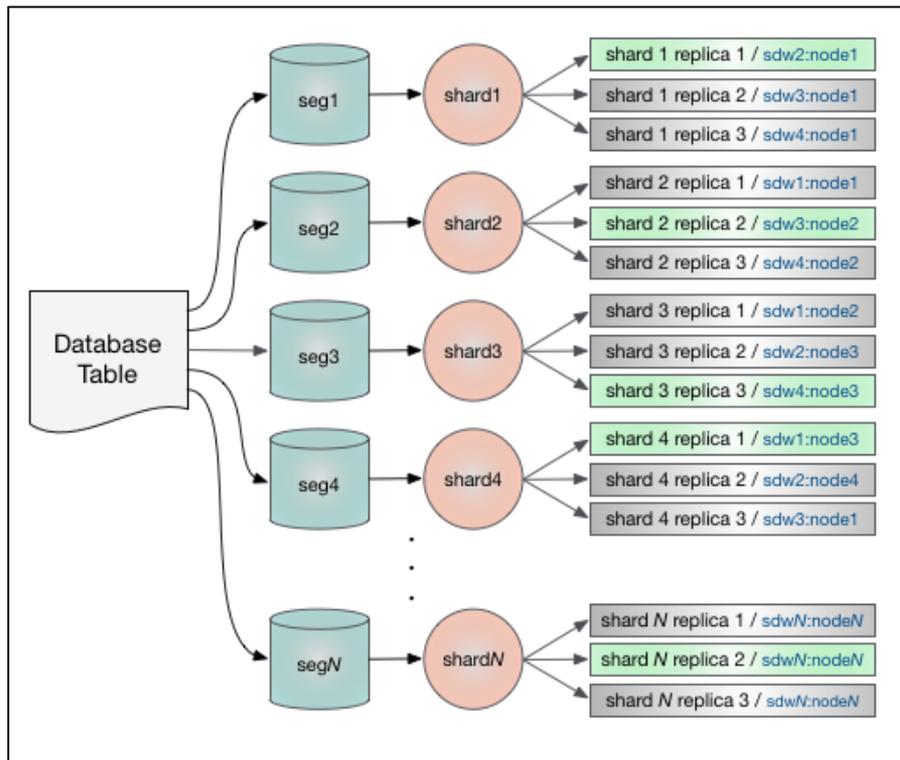
GPTText 架构简介

Greenplum 架构简介

- 用户通过master节点连接到数据库
- master节点的职责是
 - 安全认证用户连接
 - 处理 SQL 查询
 - 分配查询任务给 segments 节点
 - 搜集和呈现最终结果
- 集群内主机通过Interconnect 高效传输数据
- Segments 节点存储数据并处理查询子任务
- Shared nothing 系统架构– 集群内所有节点都有他们独立的 memory, CPU, 和 disks



GPText 高可用架构

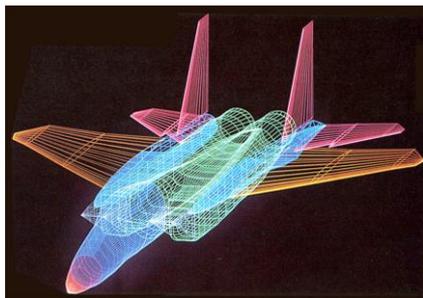
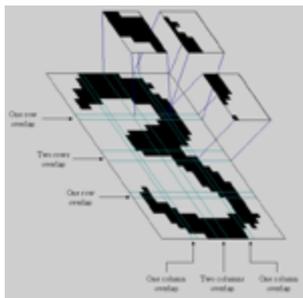


A dark, atmospheric photograph of a city street, likely in New York City, featuring classical architecture and a prominent modern skyscraper (One World Trade Center) in the background. The image is dimly lit, creating a moody and urban aesthetic. The text 'GPTText 新功能概览' is overlaid in the center in a bright white font.

GPTText 新功能概览

数据提取

很多传统企业的数据通过人能
够识别的模拟方式存储，需要
利用OCR进行数字化转换

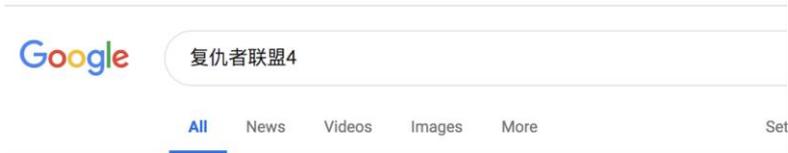


即使已经存在于计算机上的二进
制数据也会相当复杂。要提取出
来不会总是那么简单



高级检索

搜索不需要返回所有结果，搜索结果能够分页。只需要优先返回匹配度高的结果而不是106,000,000个结果



About 106,000,000 results (0.94 seconds)

复仇者联盟4：终局之战(豆瓣) - 豆瓣电影

<https://movie.douban.com/subject/26100958/?from=subject-page> [Translate this page](#)

复仇者联盟4：终局之战电影简介和剧情介绍,复仇者联盟4：终局之战影评、图片、预告片、影讯、论坛、在线购票.

复仇者联盟4：终局之战_百度百科

<https://baike.baidu.com/item/复仇者联盟4> [Translate this page](#)

《复仇者联盟4：终局之战》（Avengers: Endgame）是安东尼·罗素和乔·罗素执导的美国科幻电影，改编自美国漫威漫画，漫威电影宇宙（Marvel Cinematic Universe，缩写为MCU）第22部影片，由小罗伯特·唐尼、克里斯·埃文斯、克里斯·海姆斯沃斯、马克·鲁 ...

复仇者联盟4：终局之战 Avengers: Endgame

出品时间：2019年 拍摄日期：2017年8月10日-

[剧情简介](#) · [演员表](#) · [职员表](#) · [角色介绍](#)

复仇者联盟4：终局之战- 维基百科，自由的百科全书 - Wikipedia

<https://zh.wikipedia.org/zh-my/复仇者聯盟：終局之戰> [Translate this page](#)

《复仇者联盟4：终局之战》（英语：Avengers: Endgame）是一部于2019年上映的美国超级英雄电影，改编自漫威漫画旗下的超级英雄团队复仇者联盟，由漫威影业制作及华特迪士尼工作室电影发行。本片为2018年电影《复仇者联盟3：无限战争》的续集， ...



词性标注 POS (Part Of Speech Tagging)

词性标注 (part-of-speech tagging) ,为分词结果中的每个单词标注一个正确的词性的程序, 也即确定每个词是名词、动词、形容词或者其他词性的过程。



NER (Named Entity Recognition Tagging)

命名实体识别 (NER) 旨在将文本中的命名实体定位并分类为预先定义的类别，如人员、组织、位置、时间表达式、数量、货币值、百分比等。

Organizers: Benjamin Strauss

Bethany E. Toma

Marie de Marneffe

Alan Ritter

sportsteam India vs sportsteam Australia 2014-15 , 4th Test in geo-loc Sydney

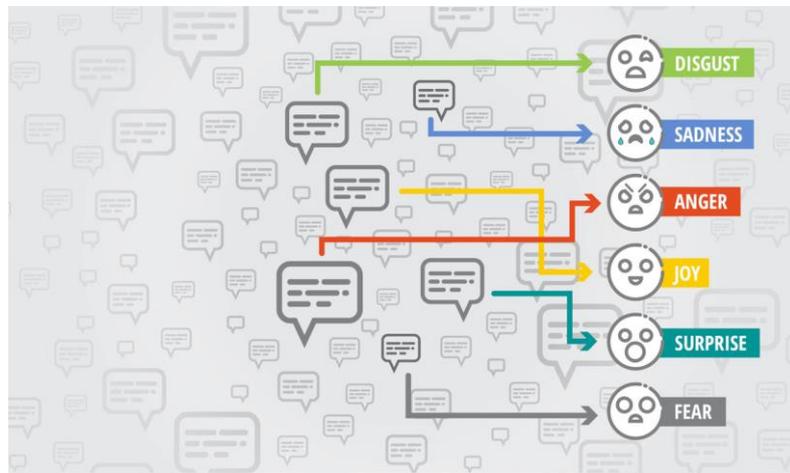
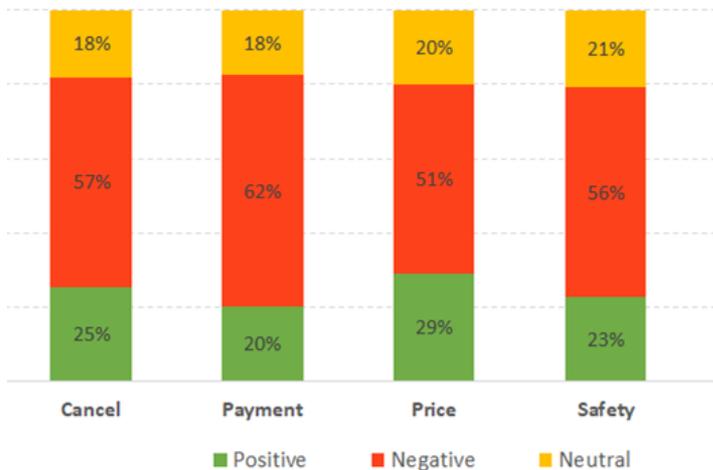
company Samsung to launch product Galaxy S6 in March

New tvshow Suits and tvshow Brooklyn Nine-Nine tomorrow ... Happy days



情感分析 (Sentiment Analysis)

文本情感分析的一个基本步骤是对文本中的某段已知文字的两极性进行分类，积极的、消极的。更高级情感分析还会寻找更复杂的情绪，比如“生气”、“悲伤”、“愤怒”

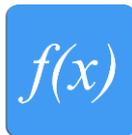


A dark, atmospheric photograph of a city street, likely in New York City, with the Freedom Tower (One World Trade Center) visible in the background. The image is dimly lit, with a blueish-grey color palette. The text 'Greenplum 大数据分析平台' is overlaid in the center in a bright white font.

Greenplum 大数据分析平台

Greenplum 数据平台中枢

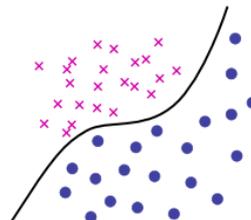
Data Transformation



Traditional BI

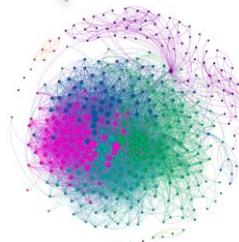
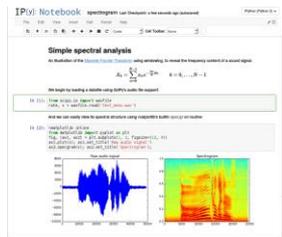


Geospatial



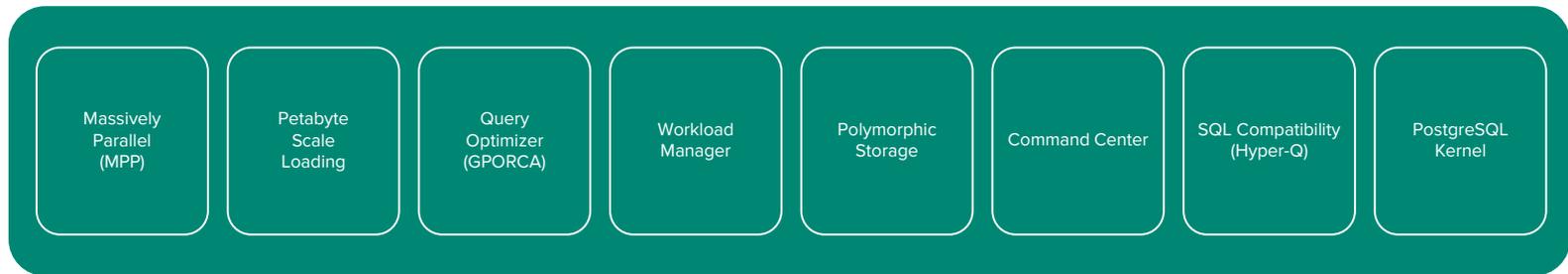
Machine Learning

Data Science
Productivity Tools



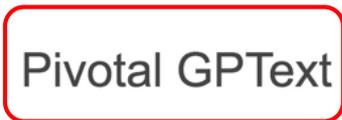
Graph

Greenplum 数据平台工具集

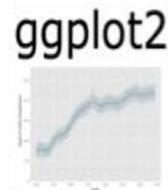


Greenplum Platform

Modeling/Analytics Tools



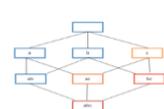
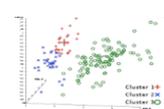
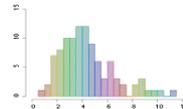
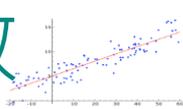
Visualization Tools



Pivotal



机器学习函数



Supervised Learning

- Neural Networks
- Support Vector Machines (SVM)
- Conditional Random Field (CRF)
- Regression Models
 - Clustered Variance
 - Cox-Proportional Hazards Regression
 - Elastic Net Regularization
 - Generalized Linear Models
 - Linear Regression
 - Logistic Regression
 - Marginal Effects
 - Multinomial Regression
 - Naïve Bayes
 - Ordinal Regression
 - Robust Variance
- Tree Methods
 - Decision Tree
 - Random Forest

Graph

- All Pairs Shortest Path (APSP)
- Breadth-First Search
- Hyperlink-Induced Topic Search (HITS)
- Average Path Length
- Closeness Centrality
- Graph Diameter
- In-Out Degree

Data Types and Transformations

- Array and Matrix Operations
- Matrix Factorization
 - Low Rank
 - Singular Value Decomposition (SVD)
- Norms and Distance Functions
- Sparse Vectors
- Encoding Categorical Variables

Comprehensive and mature data science library

Unsupervised Learning

- Association Rules (Apriori)
- Clustering (k-Means)
- Principal Component Analysis (PCA)
- Topic Modelling (Latent Dirichlet Allocation)

Nearest Neighbors

- k-Nearest Neighbors

PMML Export

- Term Frequency for Text
- Vector to Columns

Sampling

- Balanced
- Random
- Stratified

Time Series Analysis

- ARIMA

ize
ing

ics

- otive Statistics
- rdinality Estimators
- relation and Covariance
- mary

ntial Statistics

- Hypothesis Tests
- Probability Functions

Model Selection

- Cross Validation
- Prediction Metrics
- Train-Test Split

A dark, atmospheric photograph of a city street, likely in New York City, with the Freedom Tower (One World Trade Center) visible in the background. The scene is dimly lit, with a blueish-grey color palette. The text "GPTText 典型应用场景" is overlaid in the center in a bold, white, sans-serif font.

GPTText 典型应用场景

金融科技

金融风险管控:不良资产和欺诈检测

- 流式交易文档处理
- 根据交易行为和历史数据进行评估
- 法律和操作风险

客户评论数据分析

- 评论数据分类到各个不同服务部门
- 进行情感分析，分析潜在客户流失率
- 减少客户流失带来的巨大损失



生产线 - 部件质量监控

- 多个工厂数据，监控多达100多生产模型 + 部件
- 利用结构化数据 + 文本操作日志发现有瑕疵部件并发现瑕疵模型
- 利用发现的瑕疵模型找到其它类似可能有瑕疵的部件，减少部件召回率



医疗科技 - 病人风险管理

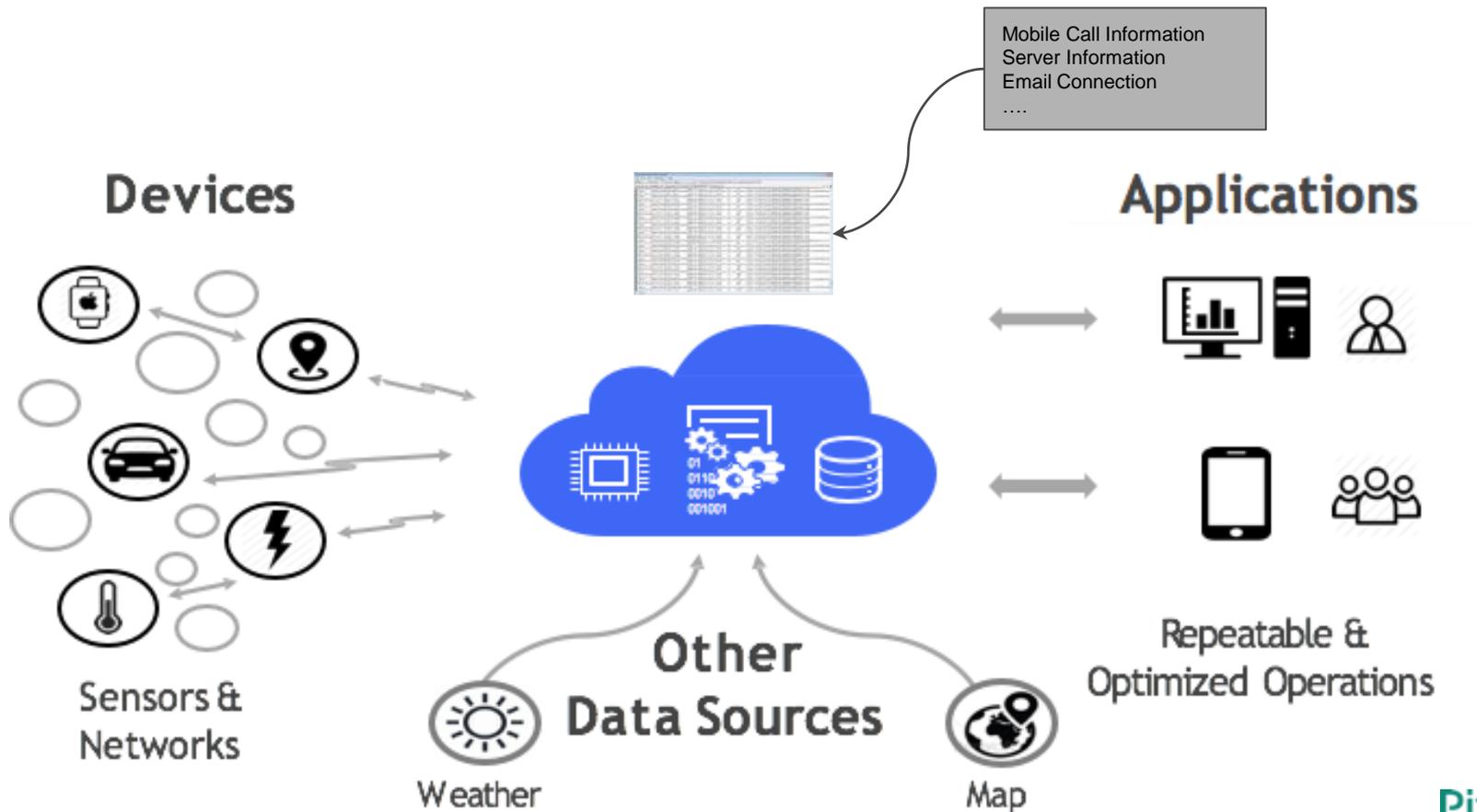
- 电子医疗记录
 - 病人指标数据(年龄, 身高, 体重, 血压, ...)
 - 医生诊断信息(自然语言形式)
- 基于数据训练模型
- 基于相近的医疗历史信息预测分析
- 辅助医生提高诊断效率



A dark, atmospheric photograph of a city street, likely in New York City, with the Freedom Tower (One World Trade Center) visible in the background. The scene is dimly lit, with a blueish-grey color palette. The text "GPTText 用于半结构化日志分析" is overlaid in the center in a bold, white, sans-serif font.

GPTText 用于半结构化日志分析

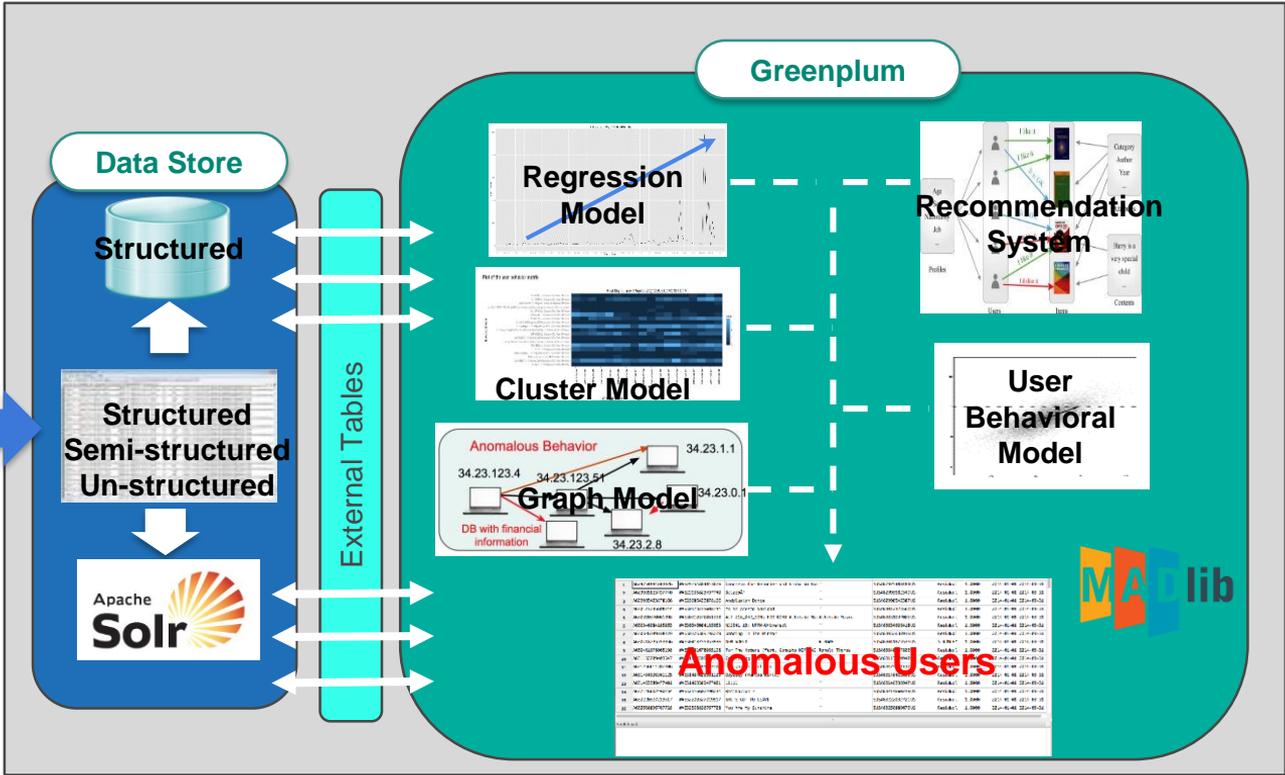
半结构化日志提取



基于日志的入侵检测分析

Logs

- Activity Record
- Email Connection
- Server Information
- Mobile Call Information



利用MADlib图算法检测异常行为

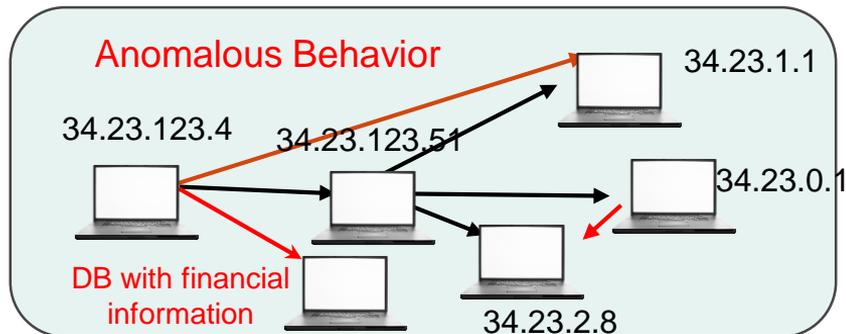
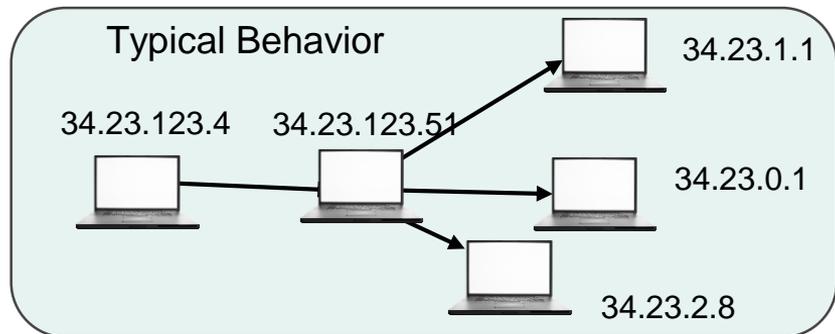
利用历史窗口事件数据为每个用户进行画像，构建用户行为图

- 用户一般从哪个机器登陆？
- 用户一般跳转到哪些机器？
- 用户登陆频率？
- 用户登陆顺序？

检测本次登陆模型是否正常？

- 用户行为是否正常？
- 是否符合某个部分用户行为模型？
- 是否符合某个岗位用户行为模型？

有向图 还是 无向图？.



GPTText 词云

A dark, low-angle photograph of a city street, likely in New York City, featuring tall buildings and a prominent skyscraper (One World Trade Center) in the center. The image is dimly lit, with a dark sky and building facades. The text "GPTText 词云" is overlaid in white, bold, sans-serif font in the center of the image.

BBC 在线文章分析

1	The city that makes the most expensive boats in the world	http://www.bbc.co.uk/news/business-40681345
2	High risk of unprecedented winter downpours - Met Office	http://www.bbc.co.uk/news/science-environment-40683302
3	How formula milk shaped the modern workplace	http://www.bbc.co.uk/news/business-40281403
4	Action urged to teach children to swim	http://www.bbc.co.uk/news/education-40685881
5	TV host's race jokes spark Brazil-Korea online war	http://www.bbc.co.uk/news/blogs-trending-40672028
6	Ryanair warns of airline fares war this summer	http://www.bbc.co.uk/news/business-40702493
7	I couldn't talk about having an eating disorder	http://www.bbc.co.uk/news/health-40681195
8	England's World Cup win: The transformation of womens cricket	http://www.bbc.co.uk/sport/cricket/40701196
9	Animal v Athlete: Four times man has raced beast	http://www.bbc.co.uk/news/world-40680346
10	How will I be flying in the future?	http://www.bbc.co.uk/guides/z3c4hv4
11	Love Jane Austen? Then find out what else you'll #LoveToRead	http://www.bbc.co.uk/programmes/articles/3W2psCW07B06Qw1kWzt2w3W/love-jane-austen-then-find-out-what-else-youll-lovetoread
12	How many selfies a day make a psychopath?	http://www.bbc.co.uk/guides/zsxgh39
...		

利用 term 计算词的重要性

词频(TF) = $\frac{\text{某个词在文章中出现的次数}}{\text{本文章总词数}}$

逆文档频率(IDF) = $\log_{10} \left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1} \right)$

TF-IDF = 词频(TF) × 逆文档频率(IDF)

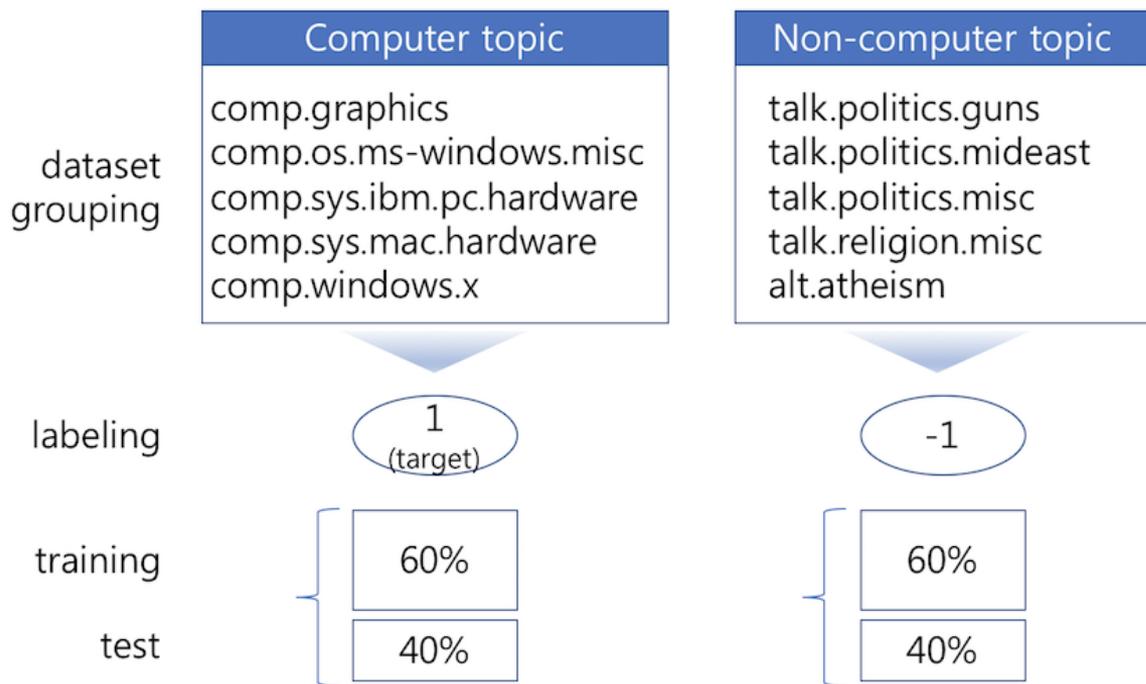
```
SELECT
    doc_id,
    bbc_articles.title,
    tf_idf,
    term
FROM (
    SELECT
        doc_id,
        tf_idf,
        term,
        rank() OVER (PARTITION BY doc_id ORDER
BY tf_idf DESC)
    FROM results2
    WHERE corpus <> 0) foo,
bbc_articles
WHERE foo.doc_id = bbc_articles.id
    AND rank <= 10
ORDER BY doc_id, rank ASC;
```


Web 网页分类

A dark, low-angle photograph of a city street, likely in New York City, featuring tall buildings and a prominent skyscraper (One World Trade Center) in the center. The text "Web 网页分类" is overlaid in white.

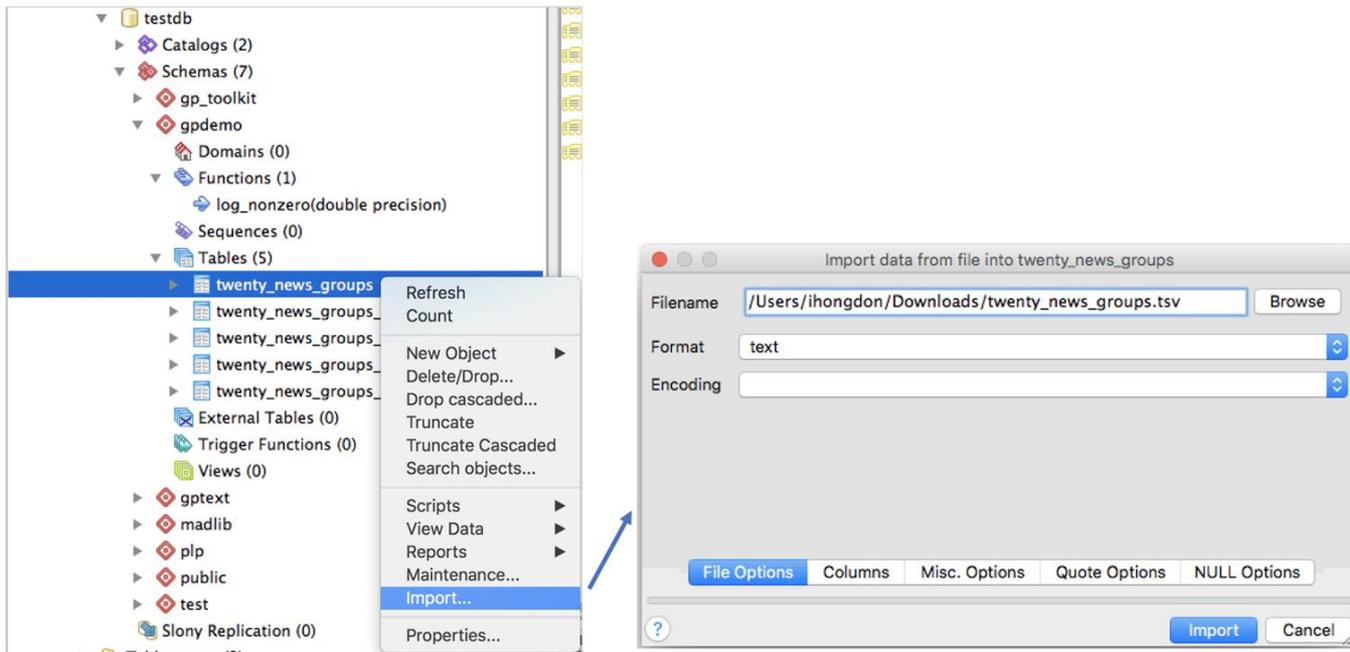
网页分类任务

Data grouping & split for training and test



网页加载到GPTText

Importing dataset using PGAdminIII's import menu



网页映射到多维空间

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\text{tf}(\text{"example"}, d_1) = \frac{0}{5} = 0$$

$$\text{tf}(\text{"example"}, d_2) = \frac{3}{7} \approx 0.429$$

$$\text{idf}(\text{"example"}, D) = \log\left(\frac{2}{1}\right) = 0.301$$

Finally,

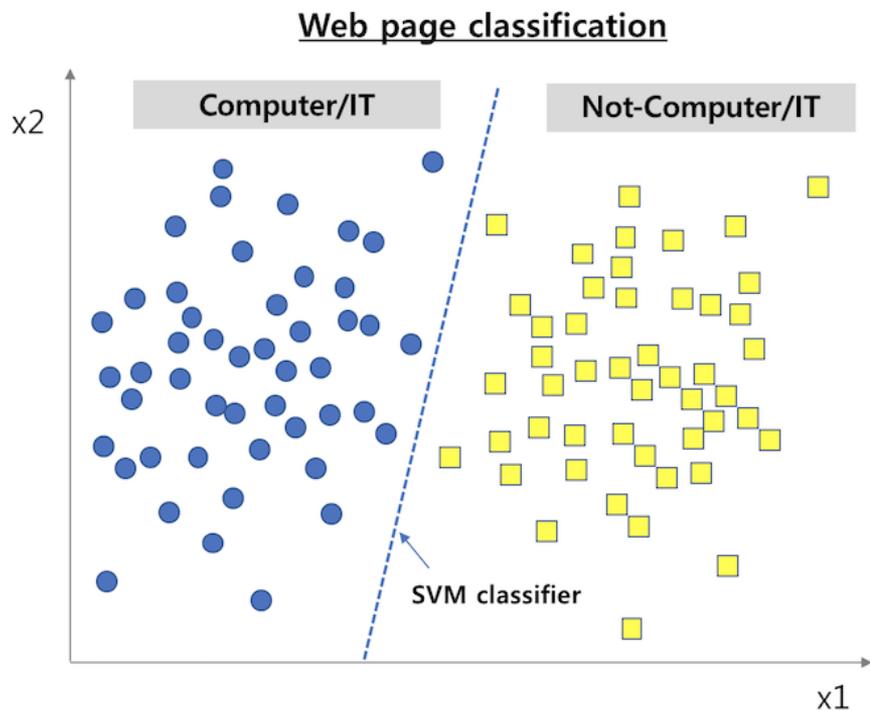
$$\text{tfidf}(\text{"example"}, d_1) = \text{tf}(\text{"example"}, d_1) \times \text{idf}(\text{"example"}, D) = 0 \times 0.301 = 0$$

$$\text{tfidf}(\text{"example"}, d_2) = \text{tf}(\text{"example"}, d_2) \times \text{idf}(\text{"example"}, D) = 0.429 \times 0.301 \approx 0.13$$

(using the [base 10 logarithm](#)).

* source: Wikipedia

利用MADlib SVM监督学习进行网页分类



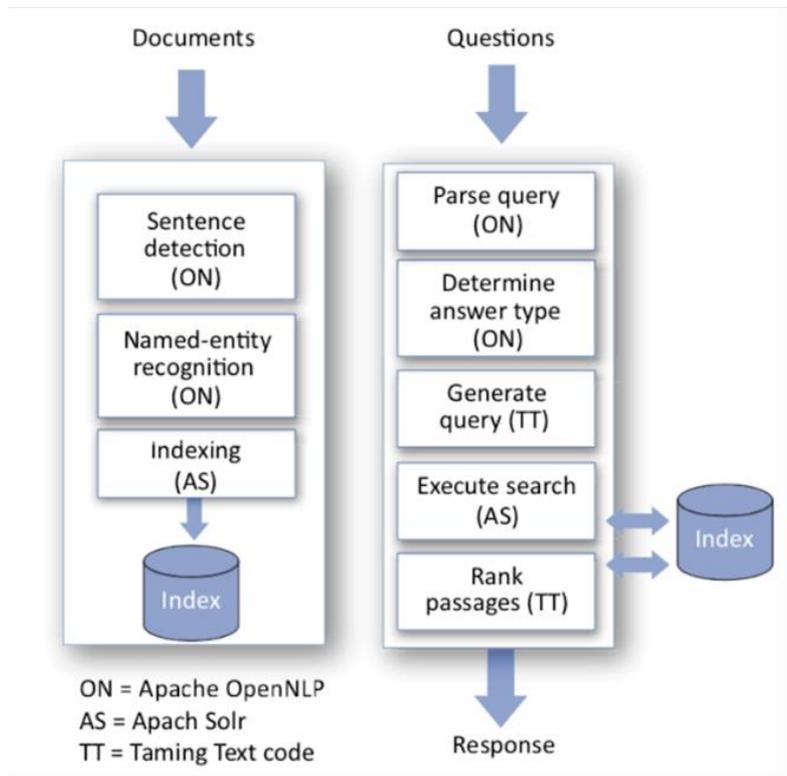
GPTText 实现问答系统

A dark, atmospheric photograph of a city street, likely in New York City, featuring tall buildings and a prominent skyscraper (One World Trade Center) in the center. The image is dimly lit, with a blueish-grey color palette. The text 'GPTText 实现问答系统' is overlaid in white, bold, sans-serif font across the middle of the image.

问答系统架构设计

- 整个系统架构分为两个部分，对文档进行POS和NER标记，和对查询问句的重写

```
<fieldType name="text_opennlp"
class="solr.TextField">
  <analyzer type="index">
<tokenizer
class="solr.OpenNLPTokenizerFactory"
...
  <analyzer type="query">
    <tokenizer
class="solr.WhitespaceTokenizerFactory"
/>
```



查询分类

需要对查询语句进行分类，标记出查询主题

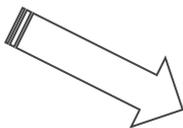
Answer type (training code)	Example
Person (P)	Which Ivy League basketball player scored the most points in a single game during the 1990s?
Location (L)	Which city generates the highest levels of sulphur dioxide in the world?
Organization (O)	Which ski resort was named the best in North America by readers of <i>Conde Nast Traveler</i> magazine?
Time point (T)	What year did the Pilgrims have their first Thanksgiving feast?
Duration (R)	How long did <i>Gunsmoke</i> run on network TV?
Money (M)	How much are Salvadoran workers paid for each \$198 Liz Claiborne jacket they sew?

查询重写

需要将原有的查询问句进行重写，转换成功能近似的查询语句

- 保留查询问句的分类标签
- 将查询问句中名词，动词等重要信息进行保留
- 词与词之间添加距离标签，比如10w

Who is US president ?



[NER_PERSON] 10w US 10w president ?



Thanks!