Semantic Search

Fast Results from Large, Foreign Language Corpora

NOVETTA

rlantz@novetta.com

Using Spark to enable Semantic Search

A low barrier to entry combined with fast and efficient distributed computing is very powerful when implementing a novel idea. Today we'll cover one such idea and why Spark was a critical component of the proof of concept.

- Introduction: What makes this hard?
- Addressing the Challenges: Why Spark?
- Semantic Search: What is the novel approach?
- Results: How does it perform?

novetta.com

Searching a foreign language corpus requires an innovative approach

- **Much of the world's knowledge is contained in documents written in foreign languages**
- **Machine translation is tricky, and word-for-word doesn't cut it**
  - This is especially true in the case of highly technical subject matter

Роль конкретных прогнозов в вымирании и восстановлении

Колби Р. Кейстлер, Эмма Хаммарлунд, Жаклин М. Баркер, Колин У. Бонд, Ральф Дж. ДиЛоне и др.

(См. Страницы 4462-4471)

Рецидив - это постоянный риск для абстинентных потребителей наркотиков, включая алкоголиков. Часто это вызвано воздействием сигналов или условий, связанных с употреблением наркотиков. Нейронные основы наркомании и рецидива были выяснены у грызунов, которые обучены нажимать рычаг для приема лекарственного средства, после чего прессование рычага гаснет, прекращая доставку лекарственного средства, а затем восстанавливается, представляя связанные с наркотиками сигналы. Такие эксперименты выявили медиальную префронтальную кору (mPFC), базолатеральную амигдалу (BLA) и ядро прилежания (NAc) как необходимое для поведения, связанного с наркотиками. Поскольку эти области имеют

中国聚合物科学学报
2017年6月，第35卷，第6期，第721-727页

由具有二苯甲基取代的芳氧基配体的半二茂钛络合物催化的1,3-丁二烯的活性聚合

Authors　Authors

博东，芮壮，张春玉，郑文杰，何新新，胡燕明，孙广平，张学全（张学全）

文件

首先在线：23 April 2017 2017年4月23日

DOI：10.1007/s10118-017-1923-8

引用本文：
Dong，B.，Zhuang，R.，Zhang，C.et al，Chin J Polym Sci（2017）35：721，doi：10.1007/s10118-017-1923-8

91
下载

抽象

制备了带有二苯甲酮取代的芳氧基配体（2a-2d）的半二茂钛配合物，其中，2c采用X射线晶体学证明的三脚扭曲四面体几何形状，通过使用这些配合物作为用甲基铝氧烷（MAO）活化的催化剂，获得了高分子量和窄分子量分布的聚-1,3-丁二烯，络合物的催化活性取决于它们的结构，具有大角度的配合物中的Ti-O-C键使它们具有更高的活性，而基于Cp *的复合物显示出比基于Cp的类似物更低的活性，络合物的活性随着聚合温度的升高而增加，而选择性保持不变，表明

Why Spark?

NOVETTA

novetta.com

Spark's ease of use and near ubiquity made it an obvious choice for analytical computations

- Our problem is to sift through potentially terabytes of unstructured foreign language data and return highly relevant results for further study and translation
- To do that we need a way to distribute the extraction and tagging computations
- Spark's ***stability***, ***availability***, and ***ease of use*** were critical to the success of this effort

The proof of the Semantic concept used Spark at almost every turn

Pre-Computation

- Entity Resolution
- Document Munging (XML to JSON)
- TF-IDF of corpus and language links
- Data validation checks

- Spark-enabled distributed computation is involved in the pre-computation, pre-search, and search phases of the engine

Pre-Search

- Scrape seed text for relevant concepts
- Build lens based on the aggregation of concepts

Search

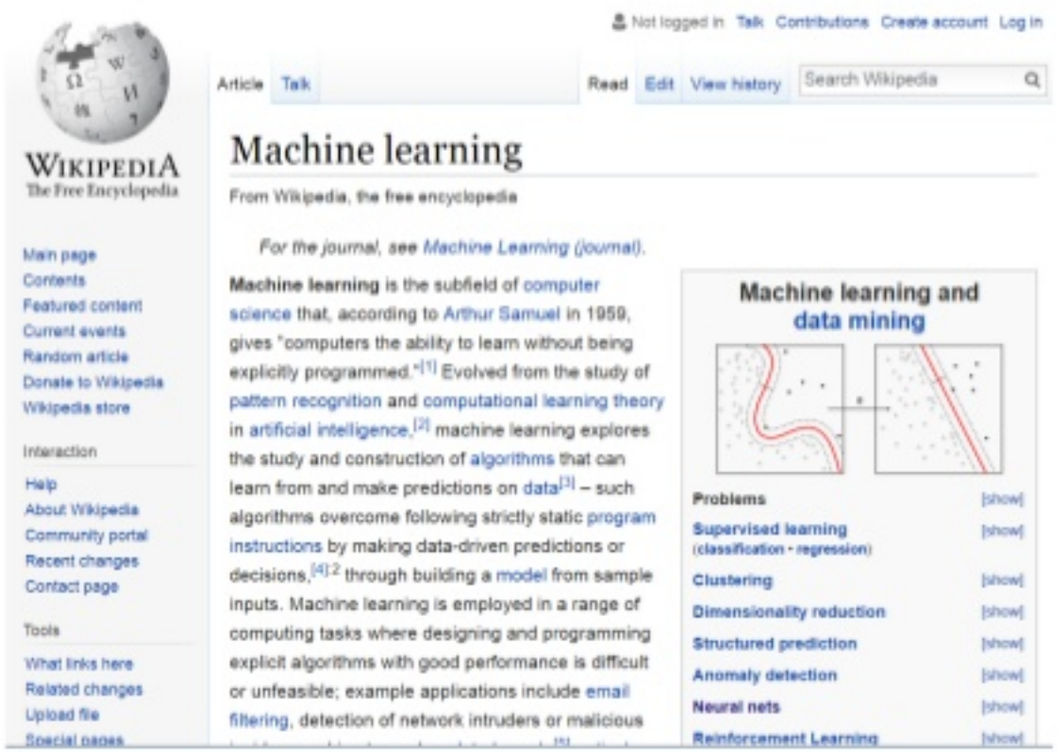
- Tag and score corpus for relevance to lens concepts
- Return ranked search results

What is Semantic Search?

novetta.com

Semantic Search is enabled by the ability to associate concepts to one another

- **The Semantic Engine takes advantage of the inter-language links that exist between topics in foreign languages in Wikipedia**



- **Spark is used to process Wikipedia articles and their foreign language counterparts, and then associate the appropriate tags to the foreign language corpus of interest**

Semantic Search can be initiated by searching directly for concepts, or scraping seed text to extract relevant concepts

English Text

Dengue is a mosquito-borne viral infection. The infection causes flu-like illness, and occasionally develops into a potentially lethal complication called severe dengue. The global incidence of dengue has grown dramatically in recent decades. About half of the world's population is now at risk. Dengue is found in tropical and sub-tropical climates worldwide, mostly in urban and semi-urban areas. Severe dengue is a leading cause of serious illness and death among children in some Asian and Latin American countries. There is no specific treatment for dengue/ severe dengue, but early detection and access to proper medical care lowers fatality rates below 1%. Dengue prevention and control depends on effective vector control measures. A dengue vaccine has been licensed by several National Regulatory Authorities for use in people 9-45 years of age living in endemic settings.

Chinese Text

登革热是一种蚊媒病毒感染。
感染导致流感样症状，有时还会发展为可能致命的并发症，称为重症登革热。
近几十年全球登革热发病率大幅度增长。
现在，约有一半世界人口面临登革热的危险。
登革热发生在全球热带和亚热带气候地带，多在城市和半城市地区。
重症登革热在亚洲和拉丁美洲一些国家是导致儿童严重患病和死亡的一个主要原因。
对登革热/重症登革热没有特异治疗办法，但及早发现和适宜的医护可将死亡率降到1%以下。
预防和控制登革热取决于有效的病媒控制措施。
一种登革热疫苗已经获得几个国家监管机构的许可，供流行区的9-45岁居民使用。

Chinese Concepts

骨痛熱症
蚊子傳播的疾病
蚊
疫苗

Tagged Concepts

Dengue Fever
Mosquito-borne Disease
Mosquito
Vaccine

Semantic Search goes beyond keywords to return meaningful results from common sense queries

- Concepts are explicitly stated, or scraped from unstructured text

- Native language concepts are compared with foreign language equivalents

- The foreign language concept is defined by the semantic meaning contained in its wiki

- Documents are scored by their concept-to-concept similarity

How does it perform?

NOVETTA

novetta.com

Using Spark enabled accelerated development, and superior computational performance

Using a cluster with 2.72 Tb of total RAM we were able to scrape and conceptualize the search text, link the concepts to their foreign language counterparts, and score the ~100 Gb corpus for relevant results in about a minute.

The ability to 'fail fast' was extremely important; allowing our data scientists and developers to experiment more freely without a fear of wasting time and resources.

Questions?

NOVETTA

novetta.com