Решение хакатона «Q&A DeepHack»

по конкурсу kaggle «The Allen AI Science Challenge»

Вихрева Мария, ВМК МГУ

29 февраля 2016

Формулировка задачи

«The Allen AI Science Challenge»

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Completed • $80,000 • 170 teams

The Allen AI Science Challenge

Wed 7 Oct 2015 – Sat 13 Feb 2016 (15 days ago)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Prizes
- Timeline

Forum

Leaderboard
- Public
- Private

My Team
- Your model
- GitHub

My Submissions

Competition Details » Get the Data » Make a submission

Is your model smarter than an 8th grader?

The Allen Institute for Artificial Intelligence (AI2) is working to improve humanity through fundamental advances in artificial intelligence. One critical but challenging problem in AI is to demonstrate the ability to consistently understand and correctly answer general questions about the world.

The Aristo project at AI2 is focused on building such a system. One way Aristo "learns" is by extracting facts from various sources and processing them into a structured

Формулировка задачи

«The Allen AI Science Challenge»

Задача – уметь отвечать на вопросы тестов для американских восьмиклассников (2500 в обучающей выборке, 800 в валидационной).

Пример

When athletes begin to exercise, their heart rates and respiration rates increase. At what level of organization does the human body coordinate these functions?

A at the tissue level

B at the organ level

C at the system level

D at the cellular level

Метрика – accuracy.

Q&A DeepHack


ОТВЕЧАЙ С ИНТЕЛЛЕКТОМ!
DeepHack.Q&A
31 Января - 6 Февраля, Москва
DeepHackLab и Московский Физико-Технический Институт
Текущие результаты
Программа
Памятка
Нон-стоп 24/7 хакатон и научная школа по самым перспективным методам построения вопросно-ответных систем и анализа текста для решения задачи The Allen AI Science Challenge
Лекции
Ведущие мировые исследователи прочитают лекции о вопросно-ответных и диалоговых системах, методах машинного перевода
Круглосуточный кодинг
Команды исследователей и программистов будут соревноваться в применении своих знаний для решения задачи

Q&A DeepHack

Команда «HoldLuceneInGloves»

Word2Vec подход

Сопоставим

- каждому слову x — вектор $WV(x) \in R^D$,
- а предложению $x_1 x_2 \dots x_n$ — нормализованную сумму векторов представлений слов $\frac{\sum_{x_i} WV(x_i)}{\left|\left|\sum_{x_i} WV(x_i)\right|\right|} \in R^D$.

Такое отображение из слов в векторное пространство умеет:

- Glove;
- обученный на собственном корпусе данных Word2Vec (библиотека gensim);
- Doc2Vec;
- ...

Word2Vec подход

Question	Answer	Схожесть(question, answer)
$x_1 x_2 \dots x_n$	$y_1^A y_2^A \dots y_{n_A}^A$	$\text{cosine}\left(\frac{\sum_{x_i} WV(x_i)}{\left\|\sum_{x_i} WV(x_i)\right\|}, \frac{\sum_{y_j^A} WV(y_j^A)}{\left\|\sum_{y_j^A} WV(y_j^A)\right\|}\right)$
	$y_1^B y_2^B \dots y_{n_B}^B$	$\text{cosine}\left(\frac{\sum_{x_i} WV(x_i)}{\left\|\sum_{x_i} WV(x_i)\right\|}, \frac{\sum_{y_j^B} WV(y_j^B)}{\left\|\sum_{y_j^B} WV(y_j^B)\right\|}\right)$
	$y_1^C y_2^C \dots y_{n_C}^C$	$\text{cosine}\left(\frac{\sum_{x_i} WV(x_i)}{\left\|\sum_{x_i} WV(x_i)\right\|}, \frac{\sum_{y_j^C} WV(y_j^C)}{\left\|\sum_{y_j^C} WV(y_j^C)\right\|}\right)$
	$y_1^D y_2^D \dots y_{n_D}^D$	$\text{cosine}\left(\frac{\sum_{x_i} WV(x_i)}{\left\|\sum_{x_i} WV(x_i)\right\|}, \frac{\sum_{y_j^D} WV(y_j^D)}{\left\|\sum_{y_j^D} WV(y_j^D)\right\|}\right)$

Ответ

$$\underset{\text{answer}\in\{A,B,C,D\}}{\arg\max} \quad \text{схожесть}(\text{question}, \text{answer})$$

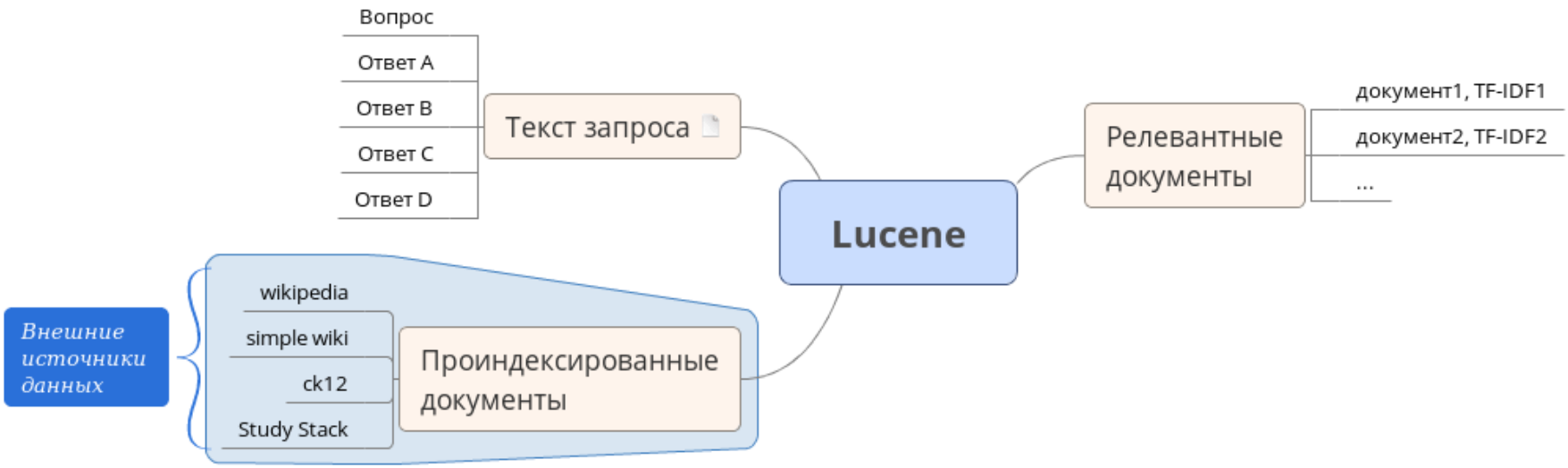
Word2Vec подход

Недостаток подхода

Word2vec-отображение не гарантирует, что вектор $\frac{\sum_{x_i} WV(x_i)}{\left|\left|\sum_{x_i} WV(x_i)\right|\right|}$ близок по косинусной метрике к представлению слова, агрегирующего смысл предложения.

Lucene(TF–IDF) подход

Lucene – библиотека для текстового поиска по корпусу данных.



Для индексации Lucene нужен корпус текста, разбитый на документы. Документами могут быть:

- статьи википедии;
- параграфы;
- предложения;
- (возможно пересекающиеся) сегменты.

Lucene(TF–IDF) подход

Оценивание ответов 1

Question	Answer	Текст запроса к Lucene	Ответ Lucene
$x_1 x_2 \dots x_n$	$y_1^A y_2^A \dots y_{n_A}^A$	$(x_1 \text{OR} x_2 \text{OR} \dots \text{OR} x_n)\text{OR}(y_1^A \text{OR} y_2^A \text{OR} \dots \text{OR} y_{n_A}^A)$	$\text{TF-IDF}_1^A, \text{TF-IDF}_2^A \dots$
	$y_1^B y_2^B \dots y_{n_B}^B$	$(x_1 \text{OR} x_2 \text{OR} \dots \text{OR} x_n)\text{OR}(y_1^B \text{OR} y_2^B \text{OR} \dots \text{OR} y_{n_B}^B)$	$\text{TF-IDF}_1^B, \text{TF-IDF}_2^B \dots$
	$y_1^C y_2^C \dots y_{n_C}^C$	$(x_1 \text{OR} x_2 \text{OR} \dots \text{OR} x_n)\text{OR}(y_1^C \text{OR} y_2^C \text{OR} \dots \text{OR} y_{n_C}^C)$	$\text{TF-IDF}_1^C, \text{TF-IDF}_2^C \dots$
	$y_1^D y_2^D \dots y_{n_D}^D$	$(x_1 \text{OR} x_2 \text{OR} \dots \text{OR} x_n)\text{OR}(y_1^D \text{OR} y_2^D \text{OR} \dots \text{OR} y_{n_D}^D)$	$\text{TF-IDF}_1^D, \text{TF-IDF}_2^D \dots$

Ответ

$$\arg\max_{\text{answer} \in \{A,B,C,D\}} \text{TF-IDF}_1^{\text{answer}}$$

или

$$\arg\max_{\text{answer} \in \{A,B,C,D\}} \sum_i \text{TF-IDF}_i^{\text{answer}}$$

Lucene(TF–IDF) подход

Оценивание ответов 2

Question	Answer	Текст запроса к Lucene	Ответ Lucene
$x_1 x_2 \dots x_n$		$x_1 \mathrm{OR} x_2 \mathrm{OR} \dots \mathrm{OR} x_n$	документ(question)
	$y_1^A y_2^A \dots y_{n_A}^A$	$y_1^A \mathrm{OR} y_2^A \mathrm{OR} \dots \mathrm{OR} y_{n_A}^A$ в документе(question)	TF-IDF_1^A, $\text{TF-IDF}_2^A \dots$
	$y_1^B y_2^B \dots y_{n_B}^B$	$y_1^B \mathrm{OR} y_2^B \mathrm{OR} \dots \mathrm{OR} y_{n_B}^B$ в документе(question)	TF-IDF_1^B, $\text{TF-IDF}_2^B \dots$
	$y_1^C y_2^C \dots y_{n_C}^C$	$y_1^C \mathrm{OR} y_2^C \mathrm{OR} \dots \mathrm{OR} y_{n_C}^C$ в документе(question)	TF-IDF_1^C, $\text{TF-IDF}_2^C \dots$
	$y_1^D y_2^D \dots y_{n_D}^D$	$y_1^D \mathrm{OR} y_2^D \mathrm{OR} \dots \mathrm{OR} y_{n_D}^D$ в документе(question)	TF-IDF_1^D, $\text{TF-IDF}_2^D \dots$

Ответ

$$\arg\max_{\text{answer} \in \{A,B,C,D\}} \text{TF-IDF}_1^{\text{answer}}$$

или

$$\arg\max_{\text{answer} \in \{A,B,C,D\}} \sum_i \text{TF-IDF}_i^{\text{answer}}$$

PMI подход

PMI (Pointwise Mutual Information) — мера схожести в теории информации и статистики.
Для пары вхождений x и y слов X и Y PMI(x, y) определяется как вероятность их совместного появления в документе в предположении независимости распределний p(x) и p(y).

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

Вероятность совместного появления слов X и Y — мат. ожидание PMI для всевозможных вариантов вхождений в документы.

PMI подход

Оценивание ответов

Question	Answer	PMI(question, answer)
$x_1, x_2, \ldots, x_n$	$y^A_1, y^A_2, \ldots, y^A_{n_A}$	$\sum_{x_i} \sum_{y^A_j} \mathrm{PMI}(x_i, y^A_j)/n_A n$
	$y^B_1, y^B_2, \ldots, y^B_{n_B}$	$\sum_{x_i} \sum_{y^B_j} \mathrm{PMI}(x_i, y^B_j)/n_B n$
	$y^C_1, y^C_2, \ldots, y^C_{n_C}$	$\sum_{x_i} \sum_{y^C_j} \mathrm{PMI}(x_i, y^C_j)/n_C n$
	$y^D_1, y^D_2, \ldots, y^D_{n_D}$	$\sum_{x_i} \sum_{y^D_j} \mathrm{PMI}(x_i, y^D_j)/n_D n$

Ответ

$$\arg\max_{\text{answer} \in \{A,B,C,D\}} \mathrm{PMI}(\text{question}, \text{answer})$$

PMI подход

Подсчет PMI(x, y)

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

Для ускорения подсчета о корпусу Википедии, PMI подсчитывался только для ТОП-1 релевантных вопросам документов по схеме:

- берем статью википедии;
- статья бьется на пересекающиеся сегменты по 100 слов;
- $p(x, y)$ — доля сегментов, в которых встретилось и слово x, и слово y;
- $p(x)$ — доля сегментов, в которых встретилось слово x;
- $p(x)$ — доля сегментов, в которых встретилось слово y.

В качестве x и y также использовались:

- биграммы;
- триграммы.

Q&A DeepHack

Completed • $80,000 • 170 teams

The Allen AI Science Challenge

Wed 7 Oct 2015 – Sat 13 Feb 2016 (2 days ago)

Dashboard

Private Leaderboard - The Allen AI Science Challenge

This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts?
Let us know.

#	Δrank	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Cardal ‡ *	0.59308	2	Mon, 08 Feb 2016 06:54:27
2	↑1	poweredByTalkwalker ‡ *	0.58344	4	Fri, 12 Feb 2016 07:28:58 (-0h)
3	↓2	Alejandro Mosquera ‡ *	0.58257	2	Sat, 06 Feb 2016 08:20:27
16	↑2	LKay ‡	0.49321	2	Sun, 07 Feb 2016 21:58:02
17	**↓1**	**HoldLuceneInGloves ‡**	**0.49102**	**10**	**Sat, 13 Feb 2016 23:22:40 (-8.8h)**
18	↑16	Balazs Godeny ‡	0.48883	2	Sat, 06 Feb 2016 08:37:08
168	–	jrdgst	0.25099	2	Sat, 13 Feb 2016 16:47:02
169	–	Bogdan Ghenea	0.25099	2	Sat, 13 Feb 2016 23:41:52
170	↓30	Čeduljko	0.23609	2	Sun, 07 Feb 2016 09:43:33

Вопросы?