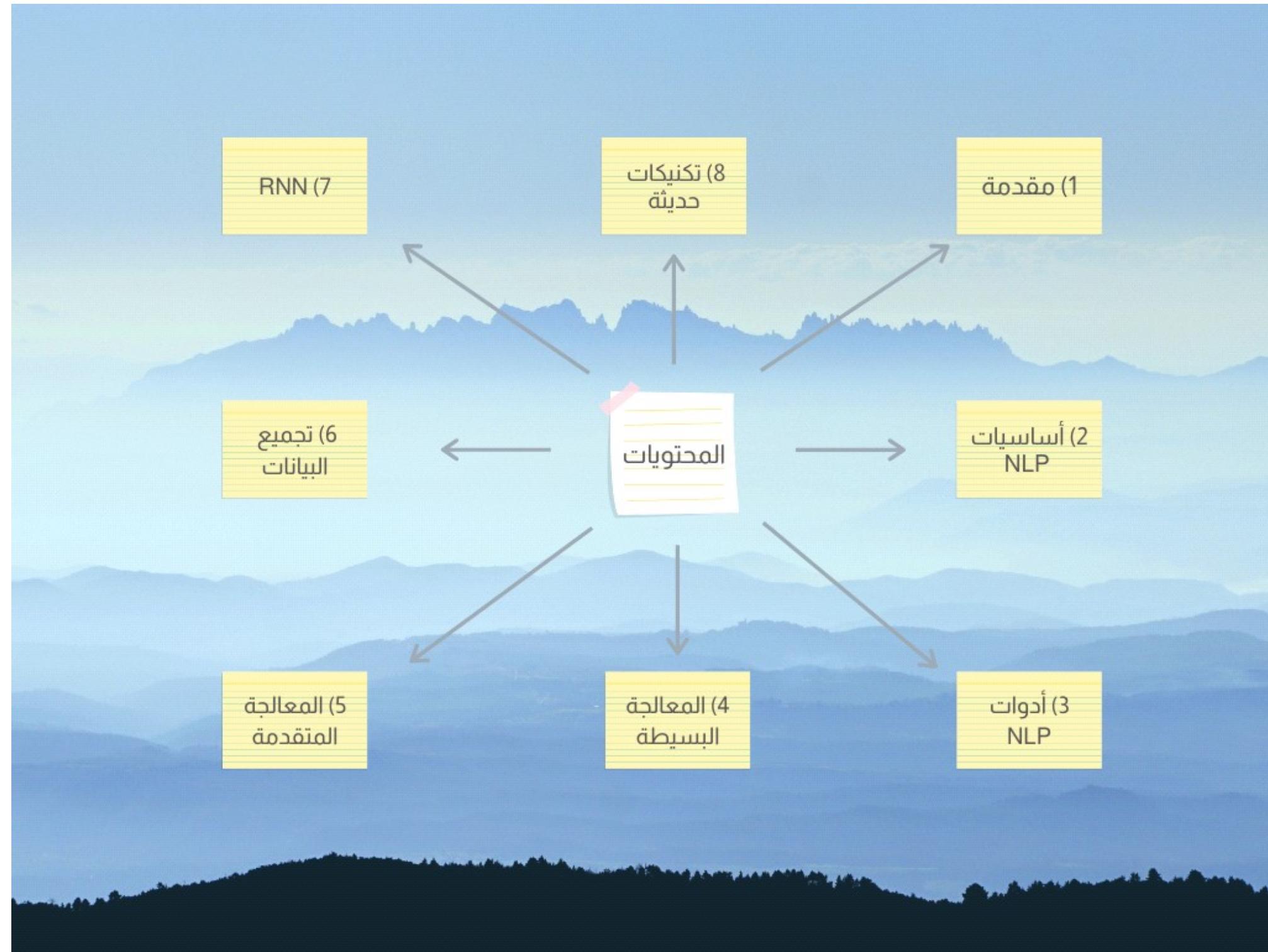


# NATURAL LANGUAGE PROCESSING

# المعالجة اللغوية الطبيعية



# المحتويات

					التطبيقات	العقبات و التحديات	NLP تاريخ ملفات pdf	ما هو NLP الملفات النصية	المحتويات المكتبات	1) مقدمة
T.Visualization	Syntactic Struc.	Matchers	Stopwords	NER	Stem & Lemm	POS	Sent. Segm.	Tokenization	NLP أساسيات	2)
	Dist. Similarity	Text Similarity	TF-IDF	BOW	Word2Vec	T. Vectors	Word embed	Word Meaning	NLP أدوات	3)
T. Generation	L. Modeling	NGrams	Lexicons	GloVe	NMF	LDA	T. Clustering	T. Classification	المعالجة البسيطة	4)
	Summarization & Snippets		Ans. Questions		Auto Correct	Vader	Naïve Bayes	Sent. Analysis	المعاجلة المتقدمة	5)
Search Engine	Relative Extraction		Information Retrieval		Information Extraction		Data Scraping	Tweet Collecting	جميع البيانات	6)
					Rec NN\TNN	GRU	LSTM	Seq to Seq	RNN (	7)
Chat Bot	Gensim	FastText	Bert	Transformer	Attention Model	T. Forcing	CNN	Word Cloud	تقنيات حديثة	8)

## القسم الخامس : المعالجة المتقدمة للنصوص

### الجزء الرابع عشر : Answering Questions

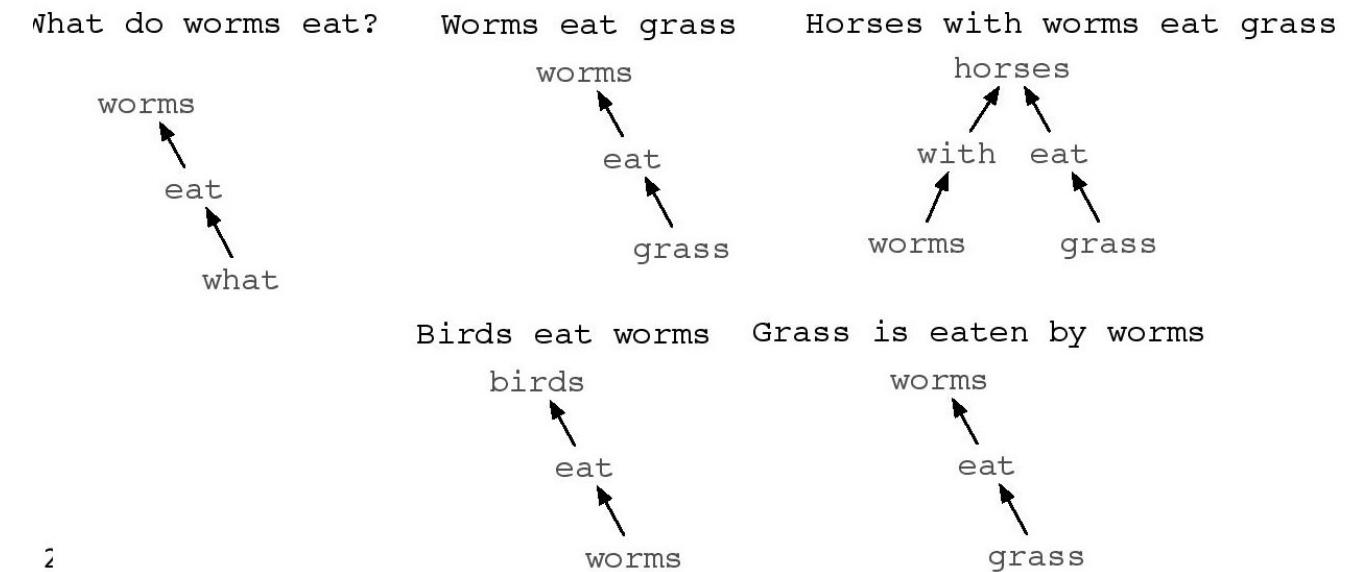
الإجابة على الأسئلة



A screenshot of the WolframAlpha search interface. The query "how many calories are in two slices of banana cream pie?" is entered in the search bar. Below the search bar, there are buttons for "Examples" and "Random". A note says "Assuming any type of pie, banana cream | Use pie, banana cream, prepared from recipe or pie, banana cream, no-bake type, prepared from mix instead". The input interpretation shows a table with rows for "pie", "amount" (2 slices), "type" (banana cream), and "total calories". The average result is 702 Cal (dietary Calories). There is a "Show details" button at the bottom right.

و هي تستخدم في العديد من التطبيقات مثل سيري او اليكسا او موقع مثل جوجل او ولفرام الفا

و هذه الفكرة قديمة ، وتم وضع تصورات عديدة لها منذ الستينات



و كانت الفكرة قائمة أولاً على تحليل السؤال لاستخراج الكلمات الهامة منه ، فلو كان لدينا سؤال :

What do worms eat ?

فتكون الكلمات الأساسية : worms , eat

ثم البحث عن العبارات التي ذكرت فيها هذه الكلمات

و هذه الطريقة بسيطة لكن ليست دقيقة ، فهي يمكن ان تأتي بإجابات سليمة مثل :

Worms eat grass

Grass is eaten by worms

او اجابات خاطئة مثل :

Birds eat worms

Horses eat grass

و تنقسم الاسئلة المستخدمة الى نوعين :

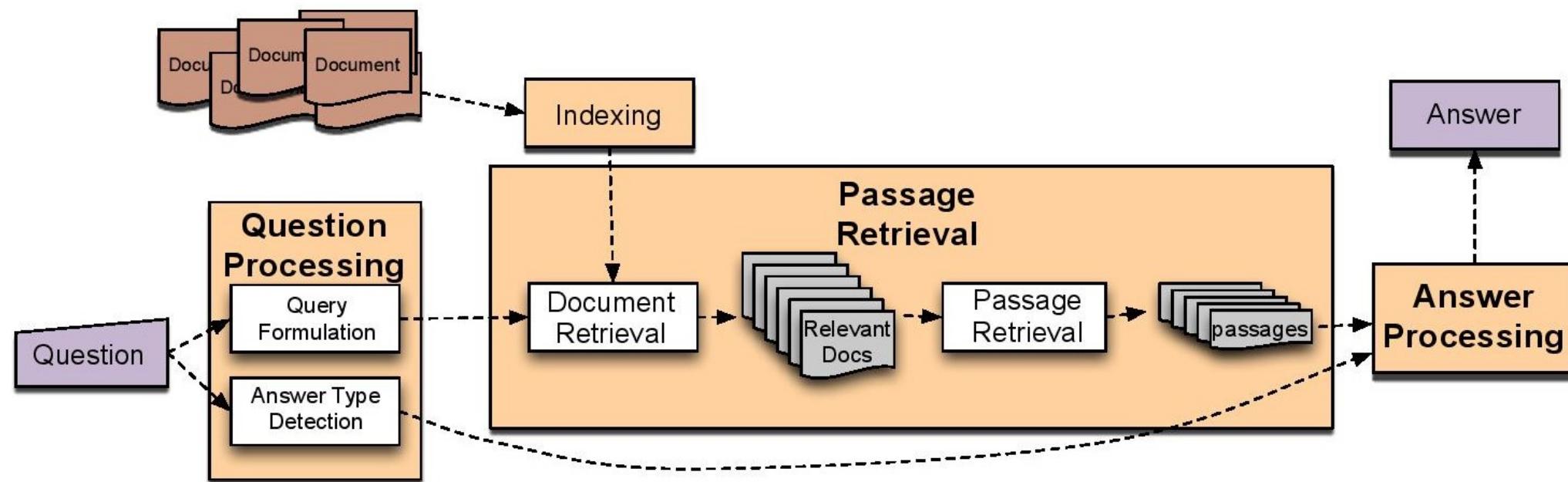
- اسئلة خاصة بالحقائق :
  - متى تم الاعلان عن حقوق الانسان
  - متى تم الاعلان عن استقلال الولايات المتحدة ؟
  - اين مقر شركة ابل ؟
  - متى ولد رخمانينوف ؟
- او اسئلة اكثر تعقيدا
  - بالنسبة للأطفال المصابين بالتوحد , تري هل يناسبهم الطعام العادي ؟ ام انه سيسبب اضرار لهم ؟
  - تري كيف تعامل الأطباء في القرون الوسطي مع وباء الطاعون دون ان يكون لديهم معدات متقدمة ؟

## و هذه عينة من الاسئلة ذات الحقائق

Question	Answer
Where is the Louvre Museum located?	In Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	The yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What is the telephone number for Stanford University?	650-723-2300

و تقوم فكرة البحث عن اجابات في خطوات متتالية :

- معالجة السؤال لاستخلاص منه شيئين :
  - كلمات البحث المطلوبة
  - نوع الاجابة المطلوب
- تناول الملفات لدينا للبحث فيها
- القيام بعمل خطوات البحث عن المعلومة IR
- تناول المعلومات المتعلقة بالأمر و ترتيبها
- معالجة الإجابة و اخراجها



كما أن التطبيقات التي تتناول الاسئلة المتعلقة بالحقائق ، يتم تخصيصها على انواع معينة من الاسئلة المتعلقة بالاوقات ،  
الاماكن ، ارقام الهاتف، قواعد بيانات المنتجات

\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*

ننناول هنا الخطوة الاولى للإجابة عن الاسئلة : و هي معالجة السؤال ، والتي تنقسم الى خطوتين :

- كلمات البحث المطلوبة
- نوع الاجابة المطلوب

و يقصد بها عدد من الأشياء مثل :

- تحديد نوع الإجابة : شخص ام مكان ام تاريخ
- استخراج كلمات البحث
- نوعية السؤال: سؤال تعريفي او رياضي او مجموعة من الاسئلة
- استخراج الكلمات ذات العلاقة بينها و بين كلمات اخري



فلا كان لدينا سؤال هكذا

They're the two states you could be reentering if you're crossing Florida's northern border

, فيتم استخراج هذه المعلومات

Answer Type: US state

Query: two states, border, Florida, north

Focus: the two states

Relations: borders(Florida, ?x, north)

و نفس الأمر مع هذه الأسئلة

*Who founded Virgin Airlines?*

- PERSON

*What Canadian city has the largest population?*

- CITY.

و بالنسبة للتقسيم ، قام هذا البحث بعمل تقسيم للإجابات المتوقعة ، في 6 محاور اساسية ، و 50 محور فرعي ، بحيث تشمل جميع الإجابات المتوقعة

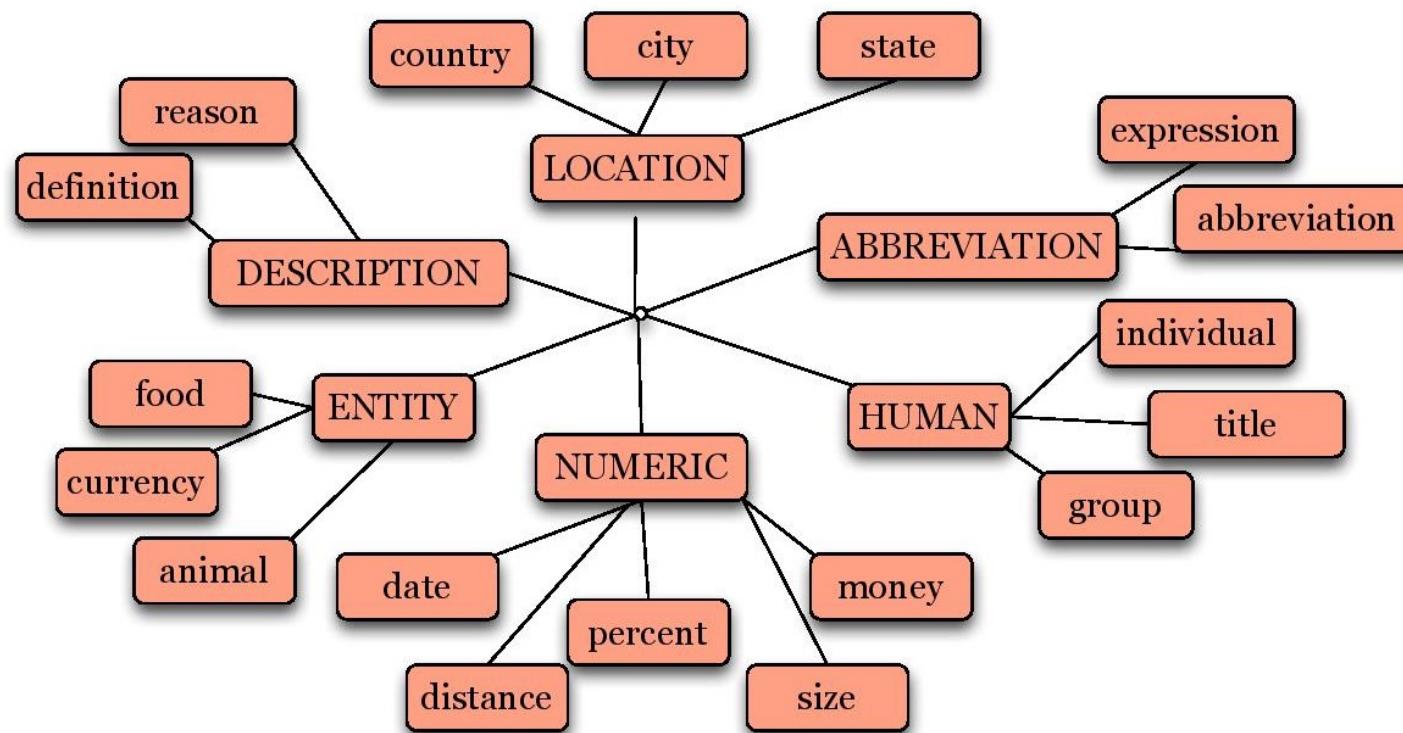
### 6 coarse classes

- ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC

### 50 finer classes

- LOCATION: city, country, mountain...
- HUMAN: group, individual, title, description
- ENTITY: animal, body, color, currency...

و هذا جزء من الشجرة



كما ان نموذج جيوباري , قام بتحليل 20 الف سؤال , ووجد ان الاجابات لها 2500 نمط

بيد أن 200 نمط منهم يغطي حوالي نصف الاجابات , وهي الانماط الاكثر شهرة , هذا هو اكثر 40 نمط منهم

- 2500 answer types in 20,000 Jeopardy question sample
- The most frequent 200 answer types cover < 50% of data
- The 40 most frequent Jeopardy answer types

he, country, city, man, film, state, she, author, group, here, company,  
president, capital, star, novel, character, woman, river, island, king,  
song, part, series, sport, singer, actor, play, team, show,  
actress, animal, presidential, composer, musical, nation,  
book, title, leader, game

\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*

## إذن كيف يتم تحديد نوع الإجابة المطلوب ؟ ؟ ؟

يتم هذا عبر ادوات عديدة ، مثل خوارزم تصنيف او اكوا德 برمجة ، فلو كان لدينا سؤال يبدأ بـ Who ( is , was , are ) ، فنعرف ان هذا سؤال عن شخص او اشخاص ، و هكذا ( were )

Regular expression-based rules can get some cases:

- Who {is|was|are|were} PERSON
- PERSON (YEAR – YEAR)

Other rules use the **question headword**:

(the headword of the first noun phrase after the wh-word)

- Which **city** in China has the largest number of foreign financial companies?
- What is the state **flower** of California?

و يتم التدريب عبر تحديد كمية كافية من الاسئلة و الاجابات ، ثم التدريب عليها هكذا

Most often, we treat the problem as machine learning classification

- **Define** a taxonomy of question types
- **Annotate** training data for each question type
- **Train** classifiers for each question class using a rich set of features.
- features include those hand-written rules!

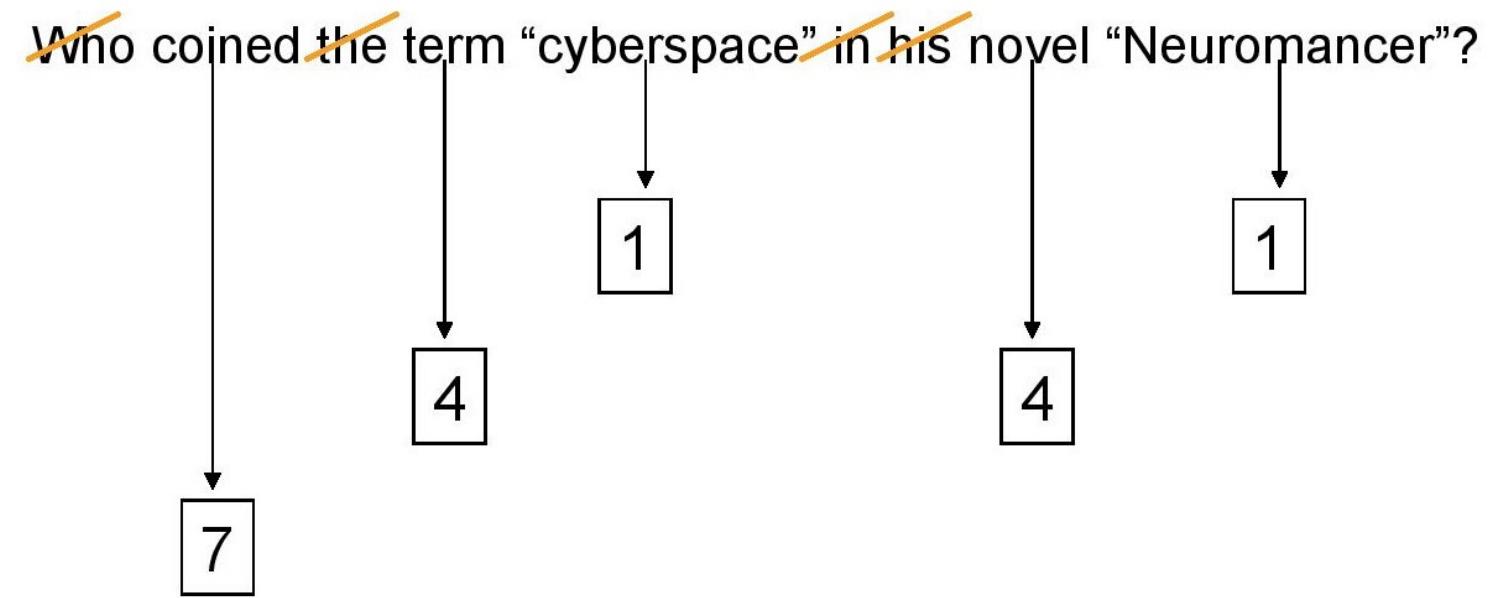
و من الفيتشرز المستخدمة :

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

### الخطوة الثانية متعلقة باستخلاص الكلمات المتعلقة بالسؤال

1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the QFW word (skipped in all previous steps)
10. Select all other words

و يتم هذا عبر عدد من القواعد هكذا



cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7

فلو كان لدينا سؤال مثل هذا , فيتم حذف stop words و انتقاء الكلمات الهمة , و عمل rank لها لمعرفة اهمية كل منها

\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*

بعد ان انتهينا من الخطوة الأولى نتناول الان الخطوة الثانية ثم الثالثة في الاجابة عن الاسئلة ، فالخطوة الثانية : retrieval اي معالجة الإجابة

و هي تقوم على خطوات ثلاث :

- البحث في الملفات بكلمات البحث IR
- عمل الإجابات و اختزالها الى فقرات أصغر لسطر واحد او عدة سطور
- عمل تقييم لكل اجابة لعرضهم بشكل مرتب

و تكون الفيترز المستخدمة هي :

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

علي أن يتم اختيار tagging للكلمات المطلوبة في الإجابة , اعتمادا علي نوع السؤال و الإجابة من الأساس  
فلو كان السؤال عن شخص ما , فنقوم بعمل tagging لأسماء الأشخاص في الإجابات لاستخراج الفيشترز , ولو كان  
السؤال عن تاريخ فيتم عمل تاج للتاريخ

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
  - Who is the prime minister of India (**PERSON**)  
**Manmohan Singh**, Prime Minister of India, had told left leaders that the deal would not be renegotiated.
  - How tall is Mt. Everest? (**LENGTH**)  
The official height of Mount Everest is **29035 feet**

لكن الأمر ليس دائماً يسير ، فيمكن أن تكون الإجابة فيها العديد من أسماء الأشخاص ، مما يجعل تحديد أيهم مهمة صعبة

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: Person

- Passage:

The Marie biscuit is named after **Marie Alexandrovna**,  
the daughter of **Czar Alexander II of Russia** and wife of  
**Alfred**, the second son of **Queen Victoria** and **Prince Albert**

و تكون هذه من الفيتشرز المستخدمة لتدريب موديل ال ML لاختيار الاجابة المناسبة

**Answer type match:** Candidate contains a phrase with the correct answer type.

**Pattern match:** Regular expression pattern matches the candidate.

**Question keywords:** # of question keywords in the candidate.

**Keyword distance:** Distance in words between the candidate and query keywords

**Novelty factor:** A word in the candidate is not in the query.

**Apposition features:** The candidate is an appositive to question terms

**Punctuation location:** The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

**Sequences of question terms:** The length of the longest sequence of question terms that occurs in the candidate answer.

و يقوم برنامج واطسن التابع لـ IBM بعمل تقييم لعدد يفوق الخمسين اجابة ، ثم اعطاء درجات لها بناء على عوامل عدّة

Each candidate answer gets scores from >50 components

- (from unstructured text, semi-structured text, triple stores)
- logical form (parse) match between question and candidate
- passage source reliability
- geospatial location
  - California is “southwest of Montana”
- temporal relationships
- taxonomic classification

و نقوم هنا بعمل تقييم للإجابات ، عبر اختيار عدد كبير عشوائي من الإجابات لأسئلة قام الخوارزم بالاجابة عليها

ثم تناول مجموعة الإجابات لسؤال معين ، والبحث بشكل يدوي أي إجابة فيهم كانت مناسبة ، وتطبيق هذا القانون

**1. Accuracy** (does answer match gold-labeled answer?)

**2. Mean Reciprocal Rank**

- For each query return a ranked list of M candidate answers.
- Query score is  $1/\text{Rank}$  of the first correct answer
  - If first answer is correct: 1
  - else if second answer is correct:  $\frac{1}{2}$
  - else if third answer is correct:  $\frac{1}{3}$ , etc.
  - Score is 0 if none of the M answers are correct
- Take the mean over all N queries

$$MRR = \frac{\sum_{i=1}^N \frac{1}{rank_i}}{N}$$

- Answers: Databases of Relations
  - born-in("Emma Goldman", "June 27 1869")
  - author-of("Cao Xue Qin", "Dream of the Red Chamber")
  - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questions: Extracting Relations in Questions

Whose granddaughter starred in E.T.?

(acted-in ?x "E.T.")

47 (granddaughter-of ?x ?y)

فمثلاً كلمة `born in` تستخدم للفصل بين اسم شخص ، و تاريخ ميلاده ، وبالتالي يتم استخدام مكتبة مثل re للبحث عن هذه الكلمة و معرفة تاريخ ميلاد شخص ما يتم السؤال عنه ، و نفس الامر في كلمات مشابهة ، وكل هذا يتم البحث عنه في ملفات ضخمة مثل ويكيبيديا و غيرها

كما ان الاسئلة المعقّدة يمكن تقسيمها لنصفين ، فلو ان لدينا سؤال : من كان القائد لفرنسا في حرب المائة عام ، فيتم البحث او لا عن التاريخ ، فنقوم بالبحث عن جملة "حرب المائة" كانت عام 1337 ، ثم جملة "الرئيس الفرنسي عام 1337 هو الملك فيليب السادس"

و يتم استخدام المعلومات ايضا لاستبعاد الاجابات غير المنطقية ، فلو تم طرح سؤال : من هو الخليفة العباسى الذى أمر بناء بغداد ؟ ، وكان لدينا اجابتين : ابو جعفر المنصور ، ام هارون الرشيد

## Relation databases

- (and obituaries, biographical dictionaries, etc.)

## IBM Watson

**“In 1594 he took a job as a tax collector in Andalusia”**

Candidates:

- Thoreau is a bad answer (born in 1817)
- Cervantes is possible (was alive in 1594)

فسيتم استبعاد هارون الرشيد ، لانه ولد عام 787 .. بينما المعلومات التاريخية تقول ان بغداد است عام 758

نفس الفكرة في المعلومات الجغرافية ، والتي حينما يكون لدينا اسماء مدن مرشحة لسؤال معين ، يمكن استبعاد اسماء مدن من المعلومات من موقع او مكتبة معينة

Beijing is a good answer for "Asian city"

California is "southwest of Montana"

geonames.org:



The screenshot shows a web browser displaying the geonames.org search results for the query "palo alto". The search bar at the top contains "palo alto" and "all countries". Below the search bar are buttons for "search", "show on map", and "[advanced search]". A message indicates "459 records found for 'palo alto'". The results are presented in a table with columns: Name, Country, Feature class, Latitude, and Longitude. The first result is "Palo Alto" (United States, California, Santa Clara County), which is also highlighted with a green background.

	Name	Country	Feature class	Latitude	Longitude
1	Palo Alto 	United States, California Santa Clara County	populated place population 64,403, elevation 9m	N 37° 26' 30"	W 122° 8' 34"
2	Palo Alto Township 	United States, Iowa Jasper County	administrative division elevation 256m	N 41° 38' 15"	W 93° 2' 57"
3	Borough of Palo Alto 	United States, Pennsylvania Schuylkill County	administrative division population 1,032, elevation 210m	N 40° 41' 21"	W 76° 10' 2"

و في حالة وجود شئ غامض ، يمكن ان يقوم الخوارزم بالرجوع للجملة السابقة ، ليعرف من يقصد بها

او ان يطرح سؤالا جديدا لتحديد الشئ المطلوب

- Coreference helps resolve ambiguities

U: “Book a table at Il Fornaio at 7:00 with **my mom**”

U: “Also send **her** an email reminder”

- Clarification questions:

U: “Chicago pizza”

S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”

\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*