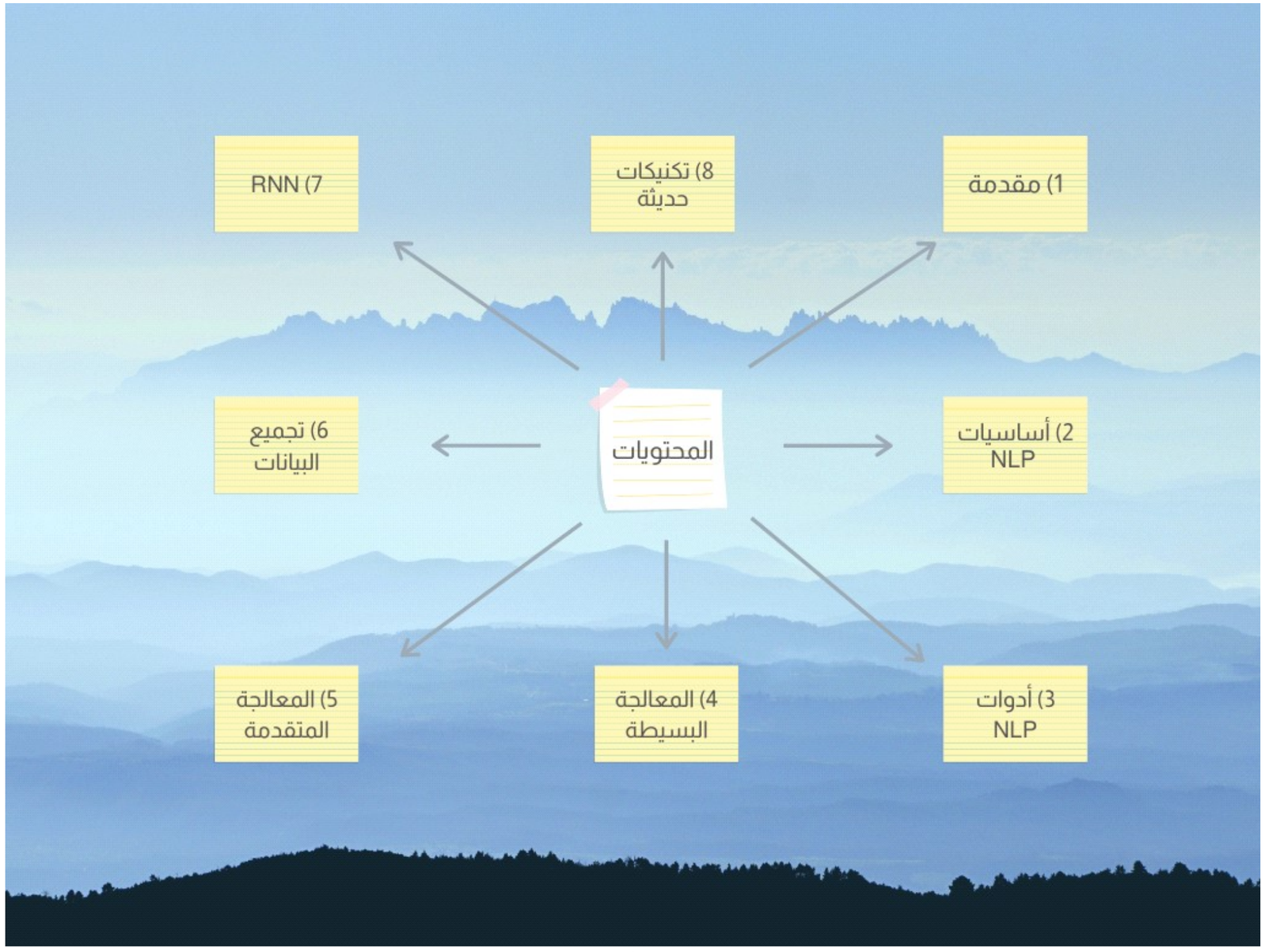
Natural Language Processing المعالجة اللغوية الطبيعية



المحتويات

القسم الخامس : المعالجة المتقدمة للنصوص

الجزء الخامس عشر : التلخيص Summarization

==========

نتناول الان أداة هامة, وهي الخاصة بتلخيص النصوص

فالهدف من هذه الخطوة : صياغة اجابة مختصرة و كافية , بحيث تكون بمعلومات كافية للمستخدم , دون استخدام جمل اكثر من اللازم

- **Goal**: produce an abridged version of a text that contains information that is important or relevant to a user.

- **Summarization Applications**
  - **outlines or abstracts** of any document, article, etc
  - **summaries** of email threads
  - **action items** from a meeting
  - **simplifying** text by compressing sentences

لها العديد من التطبيقات مثل :

- **Single-document summarization**
  - Given a single document, produce
    - abstract
    - outline
    - headline
- **Multiple-document summarization**
  - Given a group of documents, produce a gist of the content:
    - a series of news stories on the same event
    - a set of web pages about some topic or question

و هناك نوعين من التلخيص

- التلخيص من ملف واحد , و الذي من خلاله نقوم بتناول ملف واحد , واستنتاج الملخص و المقدمة
- التلخيص من ملفات عديدة , والتي فيها نتناول عدد من الملفات عن هذه النقطة , ثم نقوم بانتاج سلسلة من الاخبار او الصفحات لاحد المواقع

- **Generic summarization:**
  - Summarize the content of a document
- **Query-focused summarization:**
  - summarize a document with respect to an information need expressed in a user query.
  - a kind of complex question answering:
    - Answer a question by summarizing a document that has the information to construct the answer

ايضا هناك نوعين آخرين من التخليص , بناء علي المحتوي نفسه

- التلخيص العام : الذي يقوم بتلخيص المحتوي بشكل عام

- التلخيص المخصص , والذي يقوم بتلخيص المحتوي اعتمادا علي نقطة بحثية معينة , فيمكن مثلا تلخيص معركة ستالينجراد من ناحية معينة و هي تأثير الجو القارص علي الطرفين

_

كم ان هناك نوعين آخرين من التلخيص , بناء علي الكيفية :

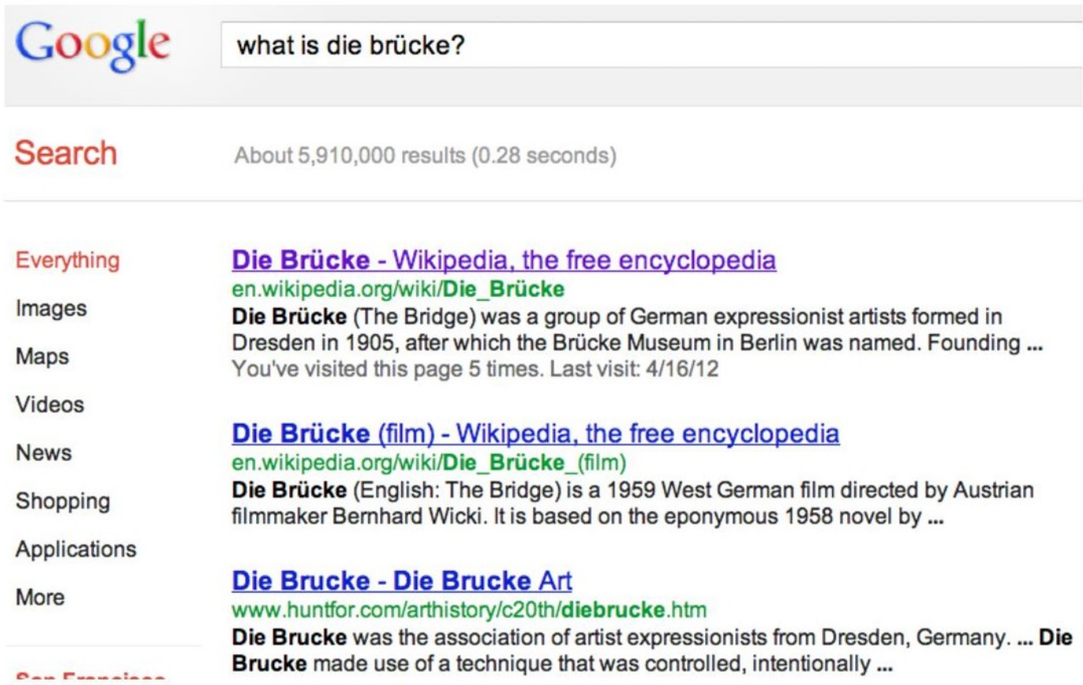
- التلخيص الاستخراجي : وهو الذي يقوم باختيار كلمات و جمل من النص الاصلي لاستخدامها كملخص للمحتوي
- التخليص الإنشائي : وهو الذي يقوم بصياغة جديدة لكلمات قد لا تكون مستخدمة في المحتوي الاصلي

و النوع الثاني اكثر كفاءة لكنه اصعب بكثير

و هناك ما يسمي snippets او المقتطفات , وهي التي تعرضها مواقع البحث تحت الرابط , لتجعل المستخدم يعرف بشكل عام الاجابة السريعة لما يريد , و هل سيدخل الموقع لان الاجابة يريدها ام هي بعيدة عنه

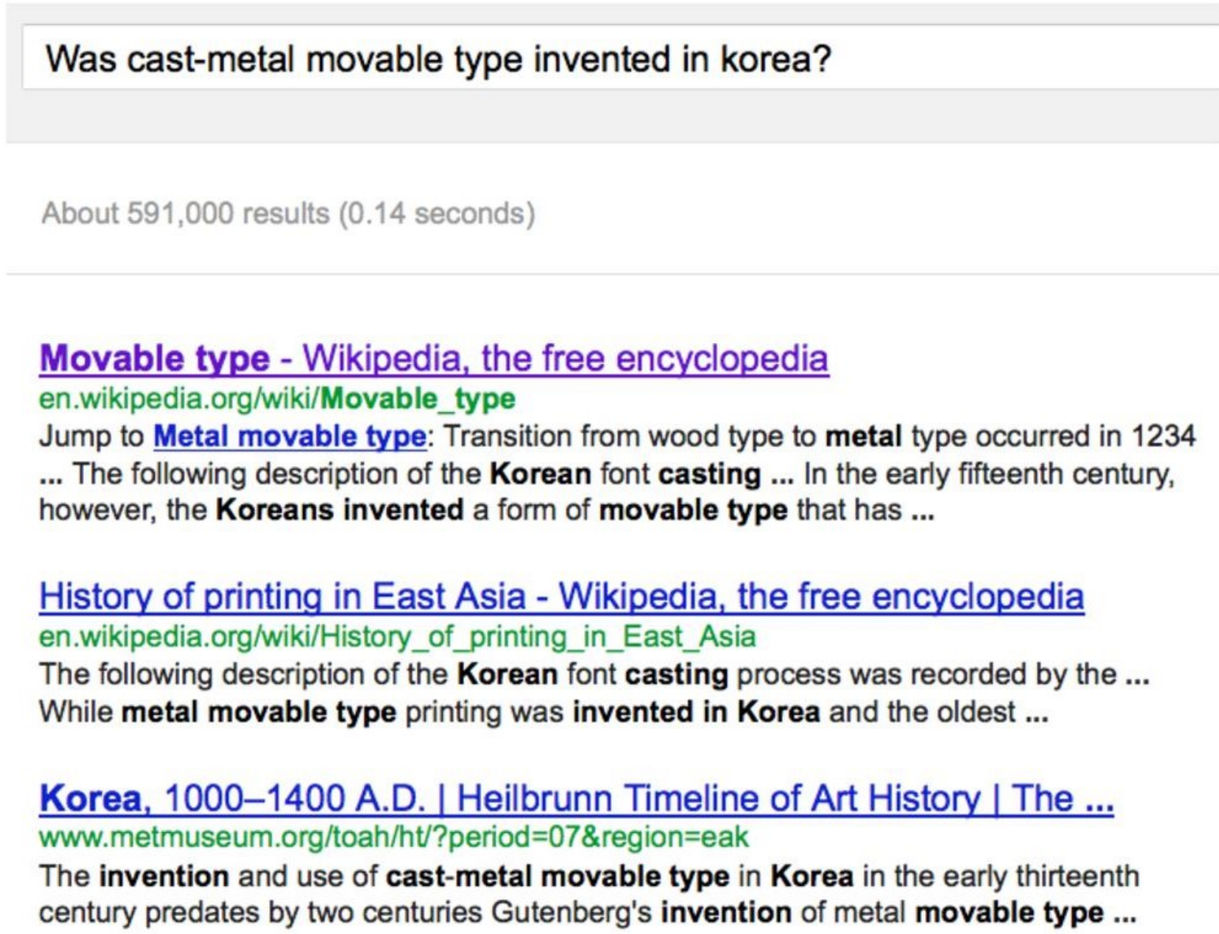
Create snippets summarizing a web page for a query

- Google: 156 characters (about 26 words) plus title and link



_

نتناول هنا كيفية انشاء هذه المقتطفات

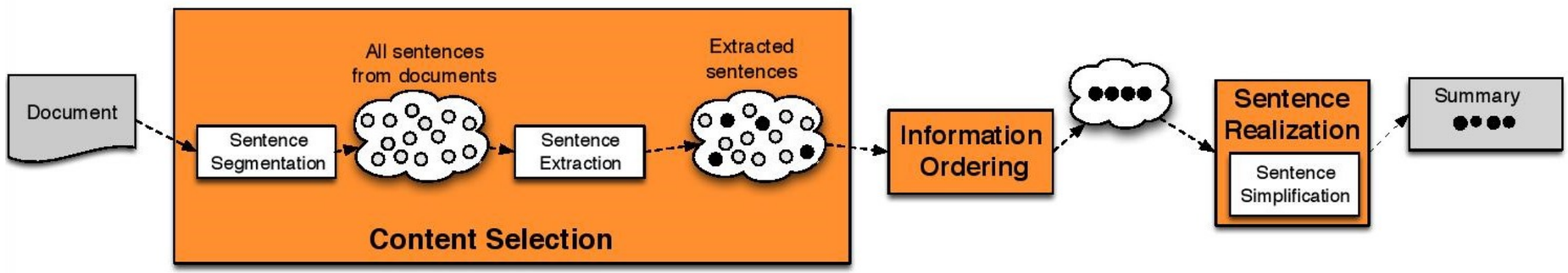


فهنا تقوم جوجل بعرض مقتطفات لكل رابط تم العثور عليه

و يتكون هذا الأمر من خطوات ثلاث :

- اختيار المحتوي : حيث نقوم بتناول المحتوي و اختيار الاجزاء او الفقرات التي تتعلق اكثر بالسؤال او كلمات البحث
- ترتيب المعلومات : عمل ترتيب معين للبيانات بحيث تكون منظمة
- ضبط العبارات , بحيث لو تقرر استخدام عدد من الجمل معا , فيتم ضبطها معا , و حذف اي كلمات زائدة

1. content selection: choose sentences to extract from the document
2. information ordering: choose an order to place them in the summary
3. sentence realization: clean up the sentences



و يتم هذا عبر طريقتين , باشراف او بدون اشراف

طريقة بدون اشراف , تقوم علي فكرة البحث عن الكلمات ذات المعلومات informative او ذات المعني الملحوظ salient

- Intuition dating back to Luhn (1958):
  - Choose sentences that have salient or informative words
- Two approaches to defining salient words
  1. tf-idf: weigh each word w_i in document j by tf-idf

$$weight(w_i) = tf_{ij} \times idf_i$$

  2. topic signature: choose a smaller set of salient words
     - mutual information
     - log-likelihood ratio (LLR) Dunning (1993), Lin and Hovy (2000)

$$weight(w_i) = \begin{cases} 1 & \text{if } -2\log\lambda(w_i) > 10 \\ 0 & \text{otherwise} \end{cases}$$

14

و هذا يتم عبر حساب الوزن باستخدام tf-idf او غيرها من المعادلات , مثل معادلة LLR الخاصة بعلاقة الكلمات بعضها البعض , اذا كانت قيمتها اكثر من 10

- choose words that are informative either
  - by log-likelihood ratio (LLR)
  - or by appearing in the query

$$weight(w_i) = \begin{cases} 1 & \text{if } -2\log\lambda(w_i) > 10 \\ 1 & \text{if } w_i \in question \\ 0 & \text{otherwise} \end{cases}$$

(could learn more complex weights)

- Weigh a sentence (or window) by weight of its words:

$$weight(s) = \frac{1}{|S|}\sum_{w \in S} weight(w)$$

15

و يتم حساب الوزن ايضا , عبر عمل معادلة , بحيث يكون لها قيمة 1 اذا كانت قيمة LLR اكبر من 10 , او اذا كانت الكلمة نفسها موجودة في السؤال .., وغير ذلك تكون بصفر

ثم يتم حساب وزن الجملة نفسها عبر ايجاد متوسط اوزان الكلمات فيها , بحيث الجملة التي تحتوي علي كلمات اكثر بوزن اعلي يكون لها افضلية

و هذه خطوات التعليم بإشراف

- Given:
  - a labeled training set of good summaries for each document
- Align:
  - the sentences in the document with sentences in the summary
- Extract features
  - position (first sentence?)
  - length of sentence
  - word informativeness, cue phrases
  - cohesion
- Train
  - a binary classifier (put sentence in summary? yes or no)

- Problems:
  - hard to get labeled training data
  - alignment difficult
  - performance not better than unsupervised algorithms
- So in practice:
  - **Unsupervised content selection is more common**

_

إذن كيف يتم تقييم هذا التلخيص ؟ ؟

يتم تقييم التلخيص عبر طريقة روج . و هي اختصار جملة : التقييم الكامل لجوهر المعني

و فكرتها تقوم علي الخطوات التالية :

- نقوم لجعل عدد من الأفراد يقومون بتلخيص الفقرة المطلوبة , لعدد محدد من الكلمات , وليكن 10 كلمات
- نقوم بجعل الخوارزم يلخص الفقرة بنفسه
- الان نقارن بين تلخيص الخوارزم و تلخيص الافراد بطريقة محددة
- ان يتم حساب عدد ال bigrams المشتركة بين تلخيص الخوارزم و تلخيص الفرد الاول , و جمعها علي عدد ال bigrams المشتركة مع الفرد الثاني و هكذا
- ثم قسمة كل هذا علي العدد الكلي لل bigrams لجميع الافراد

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Lin and Hovy 2003

Intrinsic metric for automatically evaluating summaries

- Based on BLEU (a metric used for machine translation)
- Not as good as human evaluation ("Did this answer the user's question?")
- But much more convenient

Given a document D, and an automatic summary X:

1. Have N humans produce a set of reference summaries of D
2. Run system, giving automatic summary X
3. What percentage of the bigrams from the reference summaries appear in X?

$$ROUGE-2 = \frac{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \min(count(i, X), count(i, S))}{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} count(i, S)}$$

ويمكن استخدام عدد اعلي من الـ ngrams , ولكن ستكون الكفاءة غير معبرة , ولا يمكن استخدام unigram

A ROUGE example:
Q: “What is water spinach?”

Human 1: Water spinach is a green leafy vegetable grown in the tropics.

Human 2: Water spinach is a semi-aquatic tropical plant grown as a vegetable.

Human 3: Water spinach is a commonly eaten leaf vegetable of Asia.

- System answer: Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

- ROUGE-2 = $\frac{3 + 3 + 6}{10 + 9 + 9} = 12/28 = .43$

20

فلو كان لدينا تلخيص 3 افراد , وتلخيص الخوارزم

فنري ان تلخيص الخوارزم اشترك مع تلخيص الفرد الاول في 3 bigrams و هي :
water spinach , spinach is , is a

و الثاني نفس الأمر , و الثالث اشترك معه في 6

فيكون المجموع هو 12 , و نقوم بقسمتها علي العدد الكلي لل bigrams للجمل الثلاثة و هو 28 , فتكون الكفاءة 0.43

و كلما اشترك تلخيص الخوارزم مع تلخيص الافراد اكثر , كلما زادت الكفاءة

_