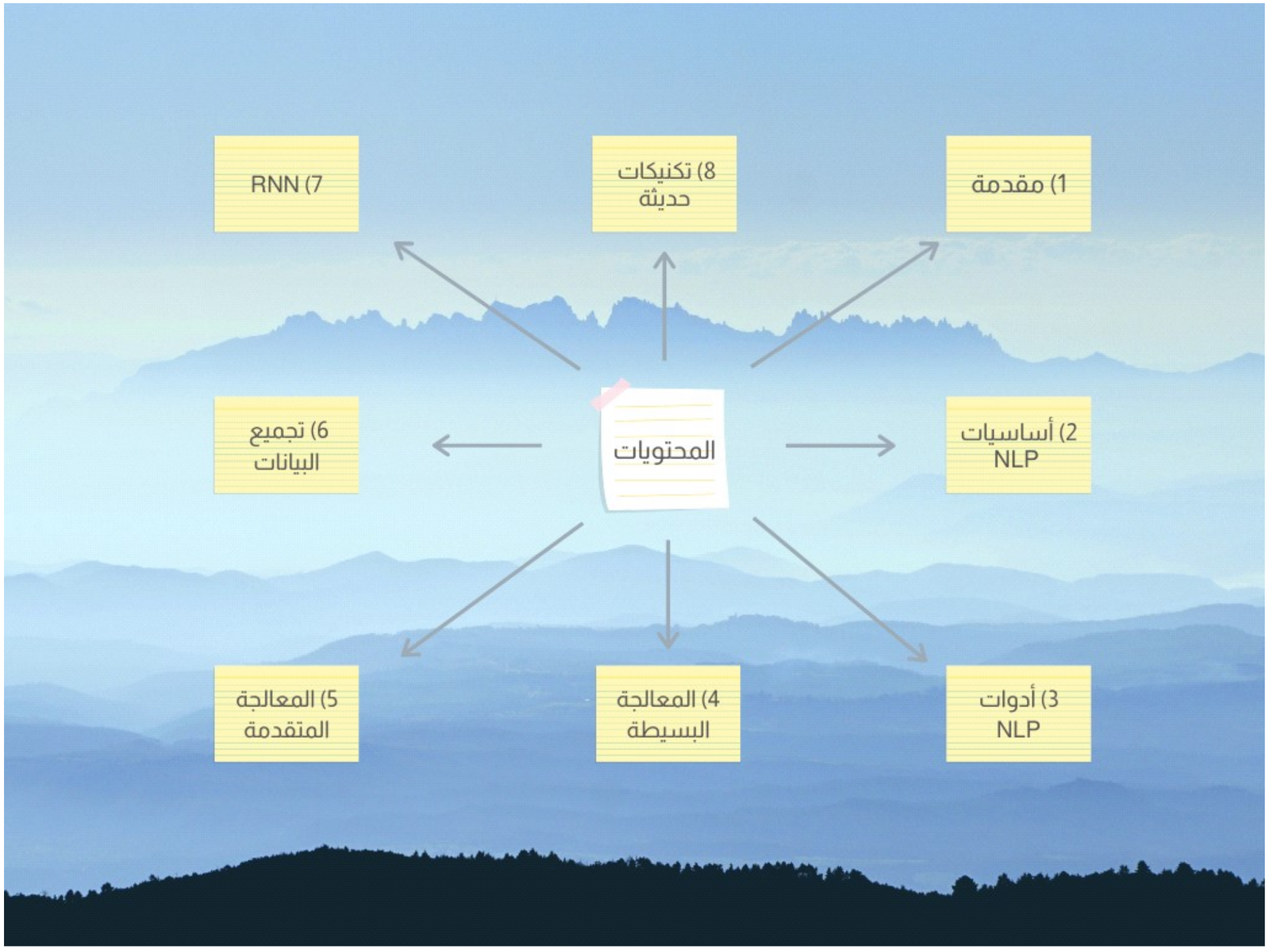
Natural Language Processing المعالجة اللغوية الطبيعية



المحتويات

القسم الرابع : المعالجة البسيطة للنصوص

الجزء السابع : Text Similarity

==========

و هو الجزء الخاص بقياس نسبة التشابه بين كلمتين , سواء في الحروف او المعني ,

و الهدف الاساسي منها هو معرفة مدي اقتراب كلمة من كلمة , وهي تستخدم بشكل أخص مع auto correct لمعرفة الكلمات القريبة من الكلمة الخطأ المكتوبة

فلو كان لدينا كلمة معينة مكتوبة خطأ مثل graffe فتكون الكلمات القريبة منها هي : giraffe , grail , graft , graf ,

و يتم هذا عبر استخدام ما يسمي أداة : minimum edit distance

و فكرة minimum edit distance تعرف علي أنها : البحث عن أقل الخطوات المطلوبة لتعديل الكلمة (الإضافة و الحذف و التعديل) لتحويل الكلمة الاولي للكلمة الثانية

فتحويل الكلمة الأصلية (الخطأ) إلي الكلمة الصحيحة , سيتم لهما عدد من خطوات ال حذف deletion و التعويض substitution و الاضافة insert حتي تتحول الكلمة الاولي للثانية

فلو كان لدينا كلمة jraaffe و مطلوب تحويلها الي الكلمة الصحيحة giraffe , فستكون هناك ثلاث خطوات

استبدال حرف j بـ g
اضافة حرف i
ازالة حرف a

* *

و هناك cost function لعملية التعديل , وذلك حتي نقوم بحساب اقل خطوات يقوم بها الخوارزم لتعديل الكلمة من الاولي للثانية ..

و يمكن ان يتم ضبط cost function بحيث تكون قيمة جميع الخطوات متشابهة, او ان يتم عمل شئ محدد اكثر , مثل ان يكون التعويض بضعف القيمة

و يتم نفس الأمر مع ترجمة الجمل الكاملة , فنقوم بمقارنة جملة مترجمة بشكل آلي , مع الجملة الحقيقية (ترجمة بشرية) لمعرفة اي كلمات سيتم اضافتها و استبدالها و حذفها

```
R Spokesman confirms    senior government adviser was shot
H Spokesman said    the senior          adviser was shot dead
              S     I             D                     I
```

علي أن الأمر ليس دائما بهذه البساطة , فيمكن أن يكون هناك جمل ذات كلمات متقاربة كثيرا , لكن المعني بعيد تماما عن التشابه

- IBM Inc. announced today
- IBM profits
- Stanford President John Hennessy announced yesterday
- for Stanford University President John Hennessy

_

ونقوم بعمل معادلة لحساب المسافة distance بين جملتين ما , و بفرض أنها بطول مختلف هكذا

و تكون الفكرة قائمة علي مقارنة اول حرف من هنا بأول حرف من هنا , ثم أول حرفين من هنا مع اول حرفين من هنا , ثم اول ثلاث حروف من هنا بأول ثلاث حروف من هنا وهكذا حتي تنتهي اقصر الكلمتين

ثم نقوم بعمل كود, بحيث يقوم بتناول كلمة كلمة , وحساب المسافة بينها و بين الكلمة الثانية , علي أن يكون للتعديل (السطر الاخير) قيمة 2

Initialization

$$D(i,0) = i$$

$$D(0,j) = j$$

Recurrence Relation:

For each $i = 1 \ldots M$

For each $j = 1 \ldots N$

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases} \end{cases}$$

Termination:

$D(N,M)$ is distance

_

علي ان هناك اوزانا علاقة الحروف مع بعضها البعض في التغير و الاستبدال , لأن هناك حروفا معرضة ان يتم استبدالها بحروف معينة أكثر من غيرها , ففي الانجليزية حرفي a , e تم الخلط بينهم , وفي العربية حرفي ـه و ـة , وهكذا

و نري هنا مصفوفة التشتت للحروف التي يتم الخلط بينها , لذا يتم استبدالها اثناء مرحلة الـ auto correct

Confusion matrix for spelling errors

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

و قد يكون السبب احيانا اقتراب حرفين من بعضهما البعض في ال keyboard مما يؤدي للخلط : (إسلام و غسلام)

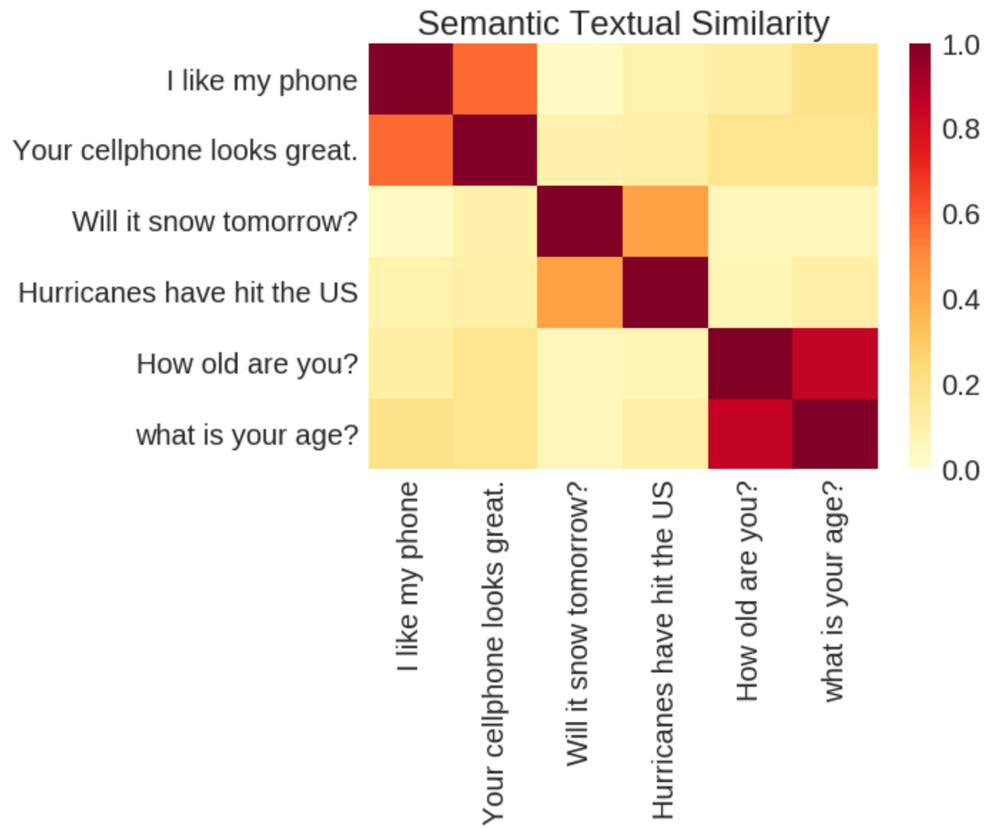
_

هناك أيضا ما يسمي المعالجة العامة للعبارات Universal Sentence Encoder

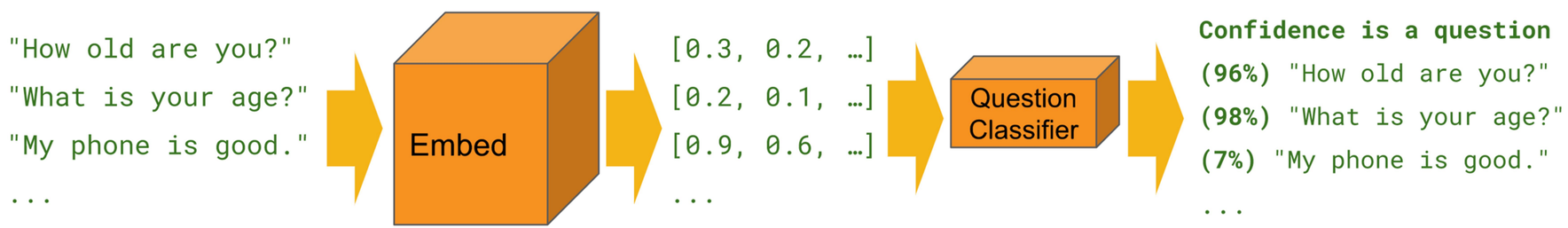
و يقصد بها التعامل مع الـ NLP بالجمل و العبارات و المقاطع الكاملة , وليس فقط بالكلمات , حيث يتم التدريب بجملة قصيرة او طويلة , او بمقطع كامل paragraph , والذي يحقق في بعض الاحيان كفاءة أعلي

و يكون المدخل بالحجم المتوقع للجمل (بعد عمل الحشو) , والمخرج يكون بحجم 512 , كما يمكن عمل semantic textual similarity, ليس هنا بين الكلمات و لكن بين الجمل و العبارات

و نري هنا ما يشبه confusion matrix لرسم العلاقة بين الجمل بعضها البعض , فيكون درجة اللون تدل علي قوة العلاقة بين كل جملة و أخري



و عبر التدريب علي عشرات الملايين من العبارات من قبل , يمكن توقع المعاني التقريبية للجمل , وكذلك يمكن تحديد اذا ما كانت جملة معينة قريبة في المعني من جملة اخري ام لا , وهل هي عبارة او سؤال



-