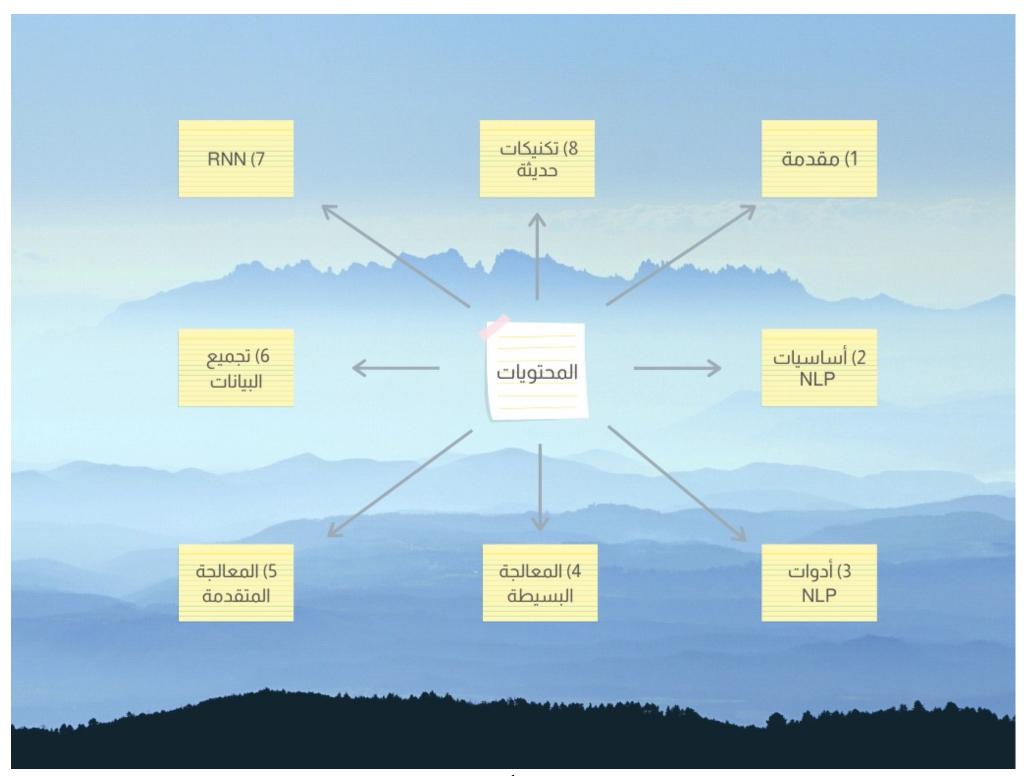
NATURAL LANGUAGE PROCESSING

المعالجة اللغوية الطبيعية



المحتويات

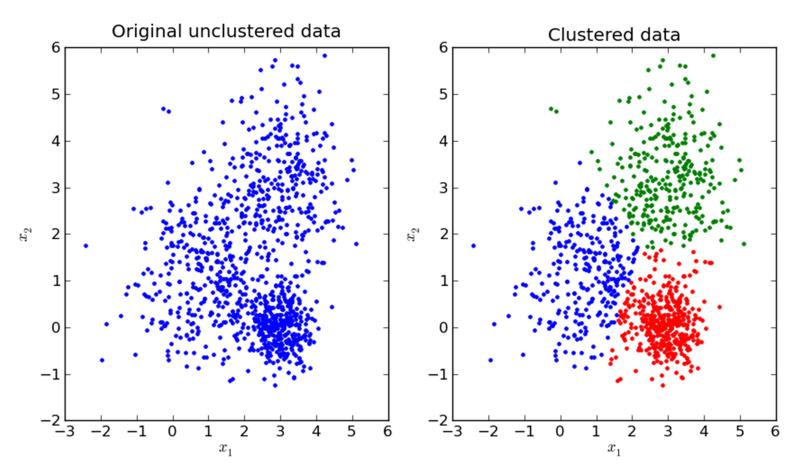
				التطبيقات	العقبات و التحديات	تاریخ NLP	ما هو NLP	المحتويات	1) مقدمة
					البحث في النصوص	ملفات pdf	الملفات النصية	المكتبات	2) أساسيات NLP
T.Visualization	Syntactic Struc.	Matchers	Stopwords	NER	Stem & Lemm	POS	Sent. Segm.	Tokenization	3) أدوات NLP
	Dist. Similarity	Text Similarity	TF-IDF	BOW	Word2Vec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

القسم الخامس: المعالجة المتقدمة للنصوص

الجزء الثاني: Text Clustering

فكرة الـ Text Clustering تعتمد علي التصنيف للنصوص غير المعنونة, أي في حالة كانت Text Clustering , و بالتالي ستكون من ادوات الـ unsupervised machine learning

و فكرة الـ clustering تعتمد علي تقسيم البيانات غير المعنونة الي فئات معينة, حسب تشابهها معا:



? ! text clustering اماذا عن الـ

لو كانت لدينا نصوص مقسمة بالفعل الي فئات واضحة : ايميلات صحيحة و سبام , نصوص مقسمة الي اقسام معينة , تعليقات ايجابية و سلبية , وقتها يمكن تطبيق احد ادوات ال text classification

بينما لو كان لدينا عدد كبير من التغريدات, وليست مقسمة الي اصناف او فئات معينة, فيمكن تقسيمها الي عناقيد clusters حيث يتم وضع كل مجموعة من التغريدات معا . .

و هناك عدد من الاستخدامات الهامة لها مثل:

- تصنيف الأخبار في وكالات الأنباء و الجهات المعنية الي اصناف محددة
 - استخلاص المعلومات من المصادر الضخمة: كتب , مقالات
 - تصنیف الـ reviews و الـ reviews
 - تصنيف الملفات و المستندات و المقالات
 - تصنيف التغريدات او التعليقات من وسائل التواصل الاجتماعي
- *_*_*_*_*_*_*_*_*_*_*_*_*_*_*

و هناك نوعين أساسيين من text clustering اما تصنيف الكلمات او تصنيف النصوص

نبدأ بتصنيف الكلمات, و التي يمكن استخدام احد خوار زميات unsupervised مع قيم ال embedding لبعض الكلمات, لتقوم بجمع الكلمات المتشابهة معا

و في حالة كان هناك عدد كبير من قيم الـ embedding , فلكي نتمكن من رسمهم , علينا او لا ان نقوم بتخفيض الابعاد , وهو ما سنستخدم فيه PCA

		d > 2	
oil	0.20		0.10
gas	2.10		3.40
city	9.30		52.1
town	6.20		34.3

How can you visualize if your representation captures these relationships?



oil & gas

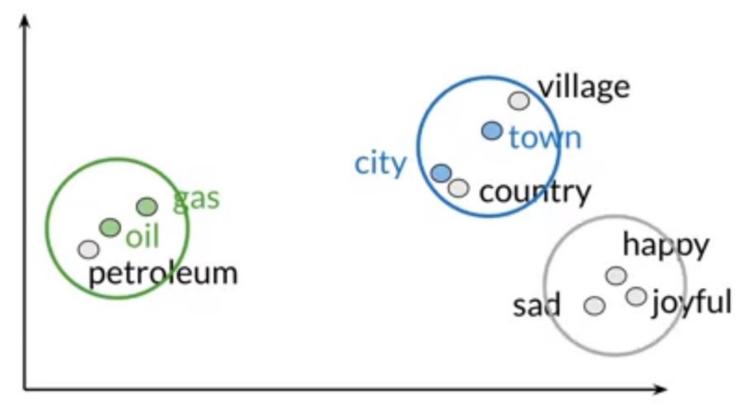


town & city

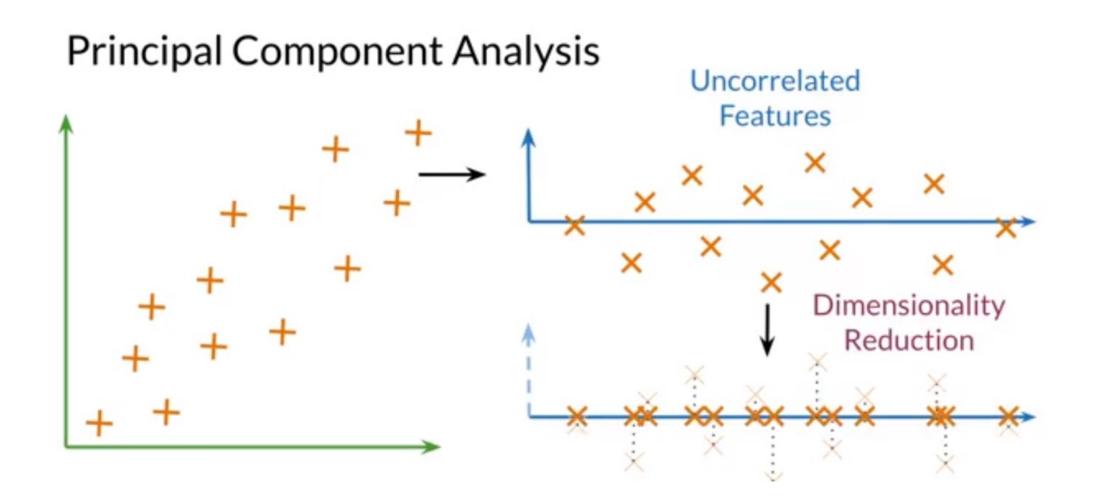
فيمكن تطبيق PCA هكذا

d > 2						d=2		
oil	0.20		0.10	_	oil	2.30	21.2	
gas	2.10		3.40	PCA	gas	1.56	19.3	
city	9.30		52.1		city	13.4	34.1	
town	6.20		34.3		town	15.6	29.8	

ثم رسمها بعد ان تحولت الي بعدين هكذا



و الية عمل PCA تعتمد علي اختيار خط ملائم, ثم اسقاط النقاط عليه بهذا الشكل



و ينجح الأمر ايضا مع اللغة العربية, في حالة تواجد قيم تضمين لكلمات اللغة العربية, وهو ما سنراه في الاقسام القادمة



ماذا عن تصنيف النصوص و العبارات ؟ ؟

يتم الأمر عبر استخراج الـ features المناسبة من النصوص, ثم ادخالها في احد خوار زميات الـ clustering

و سنري في الامثلة العملية ان ال features الاساسية هي قيم tf-idf , بينما يمكن اضافة قيم اخري , مثل :

- طول النص
- عدد الكلمات
- عدد الحروف
- تواجد الارقام
- تواجد علامات الترقيم
- تواجد كلمات معينة (كلمة اقتصادية او كلمة مسيئة على سبيل المثال)
 - غيرها من ال features التي يراها المبرمج مناسبة

كما ان هناك أداتين قويتين للـ clustering هما: الكاكم ان هناك أداتين الكاكم الك

مع العلم ان هذه الأدوات ستعمل بشكل جيد علي اللغة العربية

*_*_*_*_*_*_*_*_*_*_*_*_*_*_*