

NATURAL LANGUAGE PROCESSING

المعالجة اللغوية الطبيعية



المحتويات

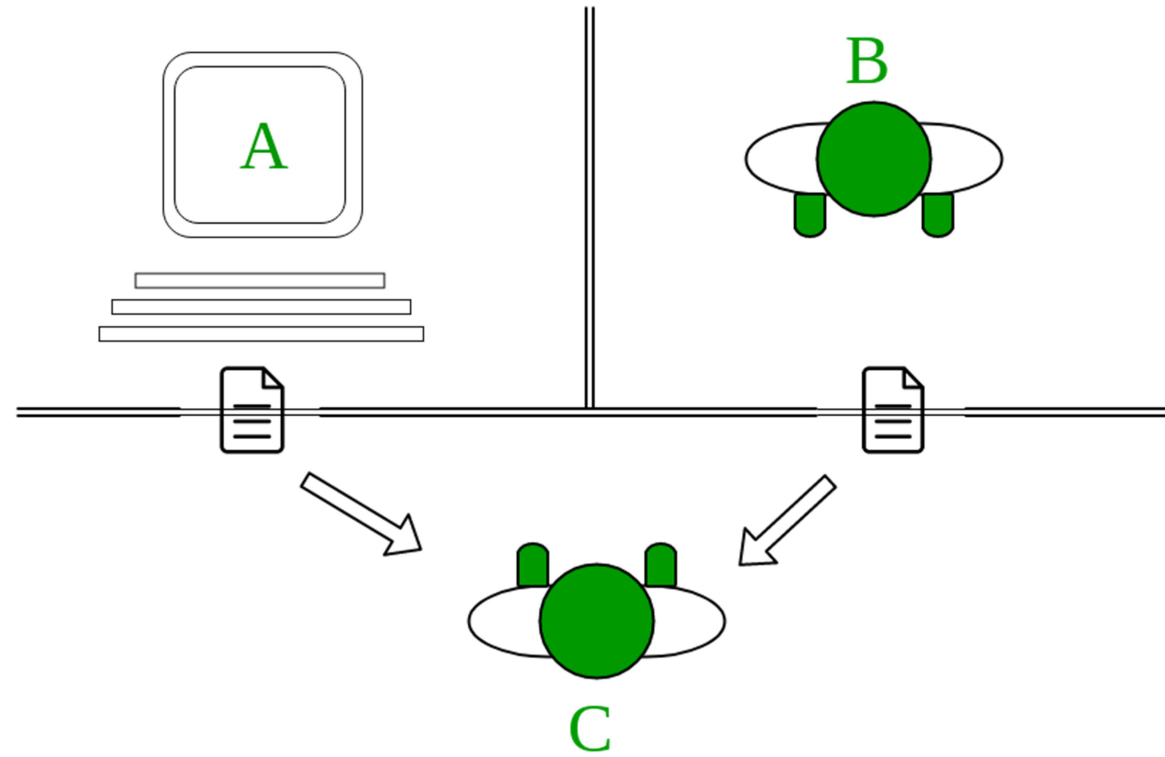
				التطبيقات	العقبات و التحديات	تاريخ NLP	ما هو NLP	المحتويات	1) مقدمة
					البحث في النصوص	ملفات pdf	الملفات النصية	المكتبات	2) أساسيات NLP
T.Visualization	Syntactic Struc.	Matchers	Stopwords	NER	Stem & Lemm	POS	Sent. Segm.	Tokenization	3) أدوات NLP
	Dist. Similarity	Text Similarity	TF-IDF	BOW	Word2Vec	T. Vectors	Word embed	Word Meaning	4) المعالجة البسيطة
T. Generation	L. Modeling	NGrams	Lexicons	GloVe	NMF	LDA	T. Clustering	T. Classification	5) المعالجة 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	7) RNN
Chat Bot	Gensim	FastText	Bert	Transformer	Attention Model	T. Forcing	CNN	Word Cloud	8) تكتيكات حديثة

القسم الخامس : المعالجة المتقدمة للنصوص

الجزء التاسع : إنتاج النصوص Text Generation

وهي من التطبيقات الهامة في NLP , حيث يتم تدريب الموديل علي عدد من النصوص المكتوبة بالفعل , بحيث يقوم فيما بعد بكتابة نصوص كاملة من تلقاء نفسه

لكن المشكلة ان اغلب خوارزميات انتاج النصوص لازالت بكفاءة محدودة , ولم تتمكن بعد من الوصول لكفاءة متميزة , او النجاح في اختبار Alan Turing



و هناك عدد من تطبيقات انتاج النصوص مثل :

- الكتابة الإبداعية
- كتابة المقالات الصحفية
- تلخيص الكتب و المقالات
- الرد علي الاسئلة
- الرد الآلي chatbot

* * * * *

ماذا عن التدريب ؟ ؟

يمكن ان يتم تدريب الخوارزم علي نصوص عامة , او نصوص محددة بفئة او نوع (اقتصادية , سياسية , أدبية) , او نصوص مخصصة بكاتب معين (أديب او كتاب صحفي)

و هنا يكون الخوارزم تم تدريبيه علي هذا النوع من النصوص , ويقوم بانتاج نصوص بكلمات و تعبيرات قريبة من النوع الذي تم تدريبيه عليه

و غالبا ما يتم الإعتماد علي فكرة ngrams , اي تحديد عدد من الكلمات السابقة لكلمة معينة , و استخدامها لتوقع الكلمة التالية

و تقوم الفكرة تدريب الخوارزم علي الخطوات التالية :

- قراءة كمية كبيرة من النصوص المحددة
- تنظيف البيانات , و ازالة اي رموز غير معروفة
- تحويل الحروف الي ارقام , حتي يتم التعامل معها في الـ RNN
- نقوم بإنتاج X و y , حيث تكون X هي عدد من الحروف المتتالية , و y الحرف التالي له
- فلو كان لدينا جملة :

I love Charles Dickens novels

و تم تحويلها الي ارقام :

50,12,3,6,87,23,6,6,54,12,25,6,3,69,85,41,23,65,22,3,14,56,98,32

فيتم اختيار اول عشر ارقام هي X و التالي لهم هو y هي :

50,12,3,6,87,23,6,6,54,12,25,6,3,69,85,41,23,65,22,3,14,56,98,32

و يتم تكرار الأمر , فيكون الارقام من الثاني الي الحادي عشر هم X رقم 12 هو y

50,12,3,6,87,23,6,6,54,12,25,6,3,69,85,41,23,65,22,3,14,56,98,32

- يتم تكرار هذا الأمر عدد من الـ epochs , حتي يتمكن الموديل من توقع اي حرف تالي
- لا تنس ان المسافة و الـ , و الـ . و الارقام و غيرها تعتبر حروف
- يقوم الموديل بتوقع حروف عشوائية معينة , ثم يقوم بادخالها في الخوارزم ليقوم باستنتاج الحرف التالي , و هكذا

* * * * *

مع التأكيد ان نفس الخوارزم يمكن استخدامه في اللغة العربية , اذا توافرت الداتا المناسبة

* * * * *