منشآت
monsha'at
الهيئة العامة للمنشآت الصغيرة والمتوسطة
Small & Medium Enterprises General Authority

مركز ذكاء

كتيّب

معالجة اللغات الطبيعية للغة العربية

إعداد وتقديم
د.دينا بنت عبد الله العباد

سياسة الاستخدام

إن المعلومات الواردة في هذا التقرير جُمِعَت ونُسِّقت بجهود موظفي مركز ذكاء، التابع لـ الهيئة العامة للمنشآت الصغيرة والمتوسطة "منشآت"، ولا ينبغي لقارئها أن يعمل بها دون مشورة مناسبة من المتخصصين.

للمزيد من المعلومات، نرجو التواصل معنا عبر البريد الإلكتروني: **support@thakaa.sa**

ماذا ستسفيد من هذا الكتيب:

- **ستتمكّن من**
  الاستيعاب والإلمام بمفاهيم معالجة اللغات الطبيعية وتقنياتها.
- **ستتعرّف على**
  بعض الأدوات والبرامج المستخدمة في مجال معالجة اللغات الطبيعية.
- **ستتمكّن من**
  تحديد تطبيق معالجة اللغات الطبيعية، واختياره بما يتناسب مع نوع المشروع ونشاط المنشأة.
- **ستتعرف على**
  خطوات تطبيق معالجة اللغات الطبيعية وإنشاء بعض التطبيقات.

المحاور:

معالجة اللغات الطبيعية (Natural Language Processing NLP):

معالجة اللغات الطبيعية (NLP) هي فرع من فروع الذكاء الاصطناعي، والذي يتيح لأجهزة الحاسب فهم اللغة البشرية المكتوبة أو المنطوقة، ومعالجتها وتوليدها.

وهو علم يجمع بين اللغويات وعلوم الحاسب الآلي.

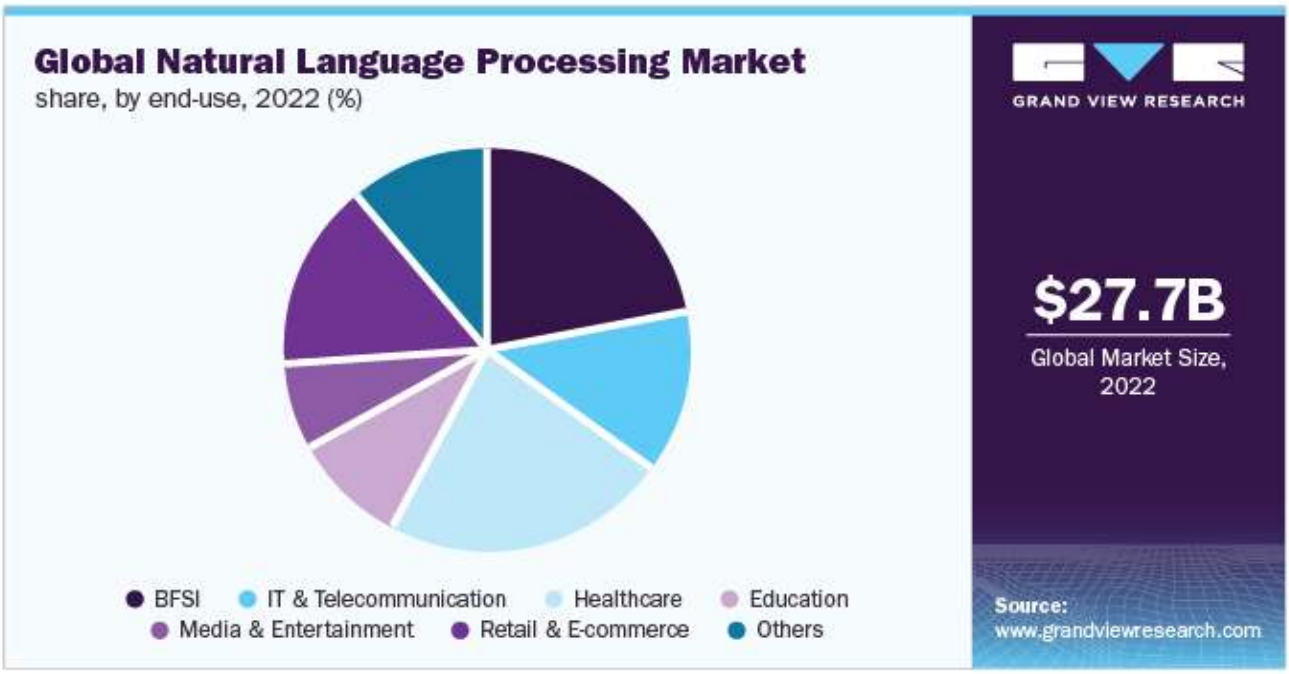
حجم السوق لمعالجات اللغات الطبيعية

حجم السوق العالمي لتطبيقات معالجة اللغات الطبيعية تم تقديره بـ 37.73 مليار دولار في عام 2022، ومن المتوقع أن يتضاعف حجم السوق بنسبة 40.4% بين 2023 و 2030.

أمثلة على شركات محلية وعالمية في مجال معالجة اللغات الطبيعية:

شركات محلية

شركات عالمية



أكثر القطاعات استثماراً في معالجات اللغات الطبيعية:

- تقنية المعلومات والاتصالات
- التجارة الإلكترونية وبيع التجزئة
- إدارة الأعمال والتسويق
- الرعاية الصحية
- التعليم
- الترفية والتواصل الاجتماعي

كُبرى الشركات المستثمرة في قطاع معالجة اللغات الطبيعية:

Apple, amazon, 3M, crayon, Google, HEALTH FIDELITY, Microsoft, ORACLE, IBM

مجالات تطبيق تقنيات معالجة اللغات الطبيعية في مجال الاعمال :


تحليل آراء العملاء


أتمتة مهام خدمة العملاء


تطبيقات المحادثة


تحليل وتصنيف السجلات والمستندات


التنبؤ بالنصوص


التدقيق اللغوي

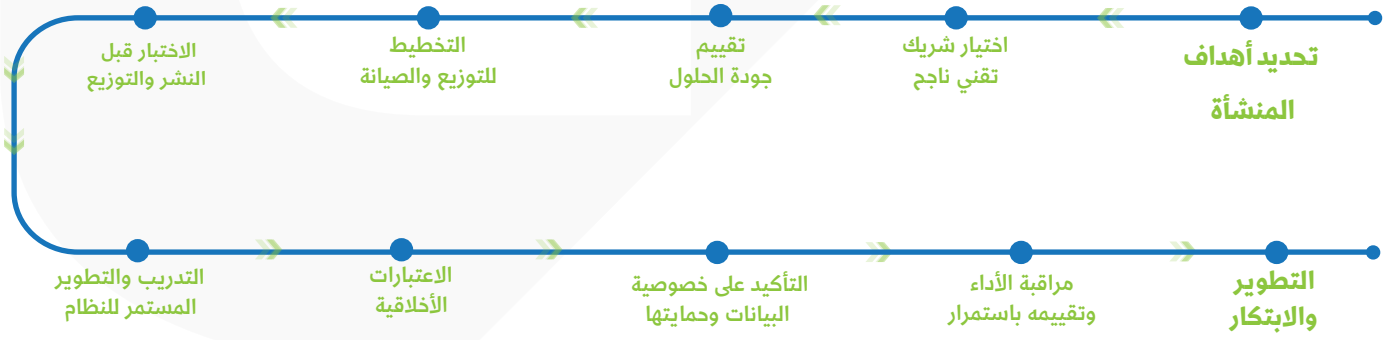

فلترة رسائل البريد الإلكتروني


كشف الانتحال

كيف ستطور تطبيقات معالجة اللغات الطبيعية من مشروعك؟

1. اتمتة العمليات الروتينية (الإجابة على استفسارات العملاء).
2. تنظيم بيانات المنشأة الضخمة وتحليلها والاستفادة منها.
3. تحسين تجربة العميل، وجودة خدمة العملاء.
4. زيادة رضا العملاء وتواصلهم.
5. تدريب الموظفين، وتقليل الأخطاء البشرية .

كيفية تحقيق الاستفادة القصوى من تطبيقات معالجة اللغات الطبيعية:



تقنيات معالجة اللغات الطبيعية واستخداماتها في قطاع الأعمال:

تحليل المشاعر Sentiment Analysis

تستخدم بشكل واسع من قبل المواقع الإلكترونية والتطبيقات، وشبكات التواصل الاجتماعية؛ لتحليل آراء الناس ومشاعرهم بخصوص خدمات الشركة أو منتجاتها:

- مراقبة مواقع التواصل الاجتماعي.
- مراقبة العلامات التجارية.
- تحليل خدمة العملاء.
- تحليل آراء العملاء .
- دراسة السوق واحتياجاته.

التعرف على الكلام (Speech recognition) وتوليد الكلام (Speech synthesis)

تُستخدم من أجل تحويل البيانات الصوتية إلى بيانات نصية، بهدف إعطاء الأوامر الصوتية، أو طرح بعض الأسئلة بشكل صوتي بدون الحاجة إلى الكتابة. تستخدم تقنية التعرف على الكلام (Speech recognition) في أنظمة المساعدة الافتراضية، وفي خدمة العملاء والاستجابة إلى شكواهم وغيرها. تقنية توليد الكلام تعمل بالاتجاه المعاكس بتحويل النص الى كلام منطوق.

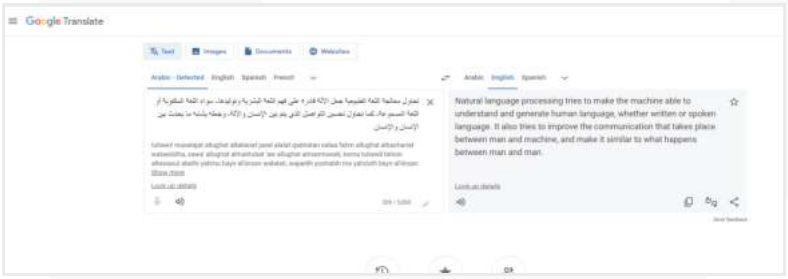
- **تصنيف الموضوعات (Topic Classification)**
- تصنيف الموضوعات حسب المحتوى (مثلا: رياضة – طب – تقنية .....)
- **تحديد النية (Intent Detection)**
- تحديد المقصد والغرض الحقيقي من الكلام (استعارة – سخرية – اقتباس...)
- **الإسناد إلى الكاتب الحقيقي (Authorship Attribution)**
- تحديد هوية الكاتب الحقيقي لنص معطى، عن طريق نمط وأسلوب الكتابة

- **استخلاص النصوص (Text Extraction)**
- استخلاص الكلمات المفتاحية (Keyword Extraction)
- استخلاص الأسماء ذات الدلالات (Named Entity Recognition) أو (NER)

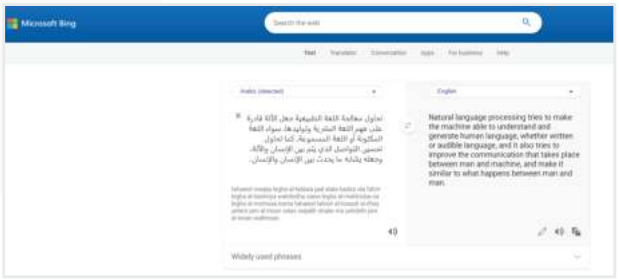
- **الترجمة الآلية (Machine Translation)**
- **تلخيص النصوص وإعادة كتابتها (Text Summarization and rewriting)**

الترجمة الآلية (Machine Translation)

قيام برامج الحاسب بترجمة النص المدخل إلى لغات مختلفة.



Google Translate



Smartcat

تلخيص النصوص وإعادة كتابتها (Text Summarization and rewriting)

مُلخِّص النص هو أداة إلكترونية تستخدم الذكاء الاصطناعي والخوارزميات المعقدة؛ لتكثيف النص من نسخته الطويلة والمفصلة، إلى نسخة قصيرة وسهلة الفهم.

Arabic Summarizing Tool FREE (paraphrasetool.com)

تطبيقات المحادثة الآلية Chatbot

هي برامج تحاكي المحادثة البشرية (المكتوبة أو المنطوقة) وتعالجها، ما يتيح للبشر التفاعل مع الأجهزة الرقمية، كما لو كانوا يتواصلون مع شخص حقيقي.

Bots Listing | arabot Bot Builder

تقنيات معالجة اللغات الطبيعية

- التعرف على الصوتيات
- التعرف على أنماط الكلام
- التحليل الصرفي
- التحليل النحوي
- التحليل الدلالي
- تصنيف النصوص
- فهم النصوص
- التشكيل الآلي
- توليد الكلام من النصوص
- توليد النصوص
- إملاء النصوص
- القراءة الآلية للنصوص
- تمييز الكلام
- التدقيق الصرفي
- التدقيق الإملائي
- تنقيح النصوص
- الترجمة الآلية
- فهم الأسئلة والإجابة عليها
- استرجاع المعلومات
- استخلاص المعلومات
- التلخيص التلقائي
- التنقيب في النصوص
- البحث عن المعلومات
- نظم التعليم الذكية

تحديات معالجة اللغات الطبيعية للغة العربية

كم عدد اللهجات - عدد أشكال اللغة العربية ولهجاتها.
في اللغة العربية؟ كم لهجة في المملكة فقط؟

- التشكيل (الحركات)

مثلا: عبارة (درس الطالب) ماذا تعني؟
هل تعني بأن الطالب قام بعملية الدراسة؟
أم أن الطالب تم تدريسه(دُرِّس الطالب)
أم أن الطالب تمت دراسته كحالة (دُرِسَ الطالب)
التشكيل يغير في معنى الجملة كلياً

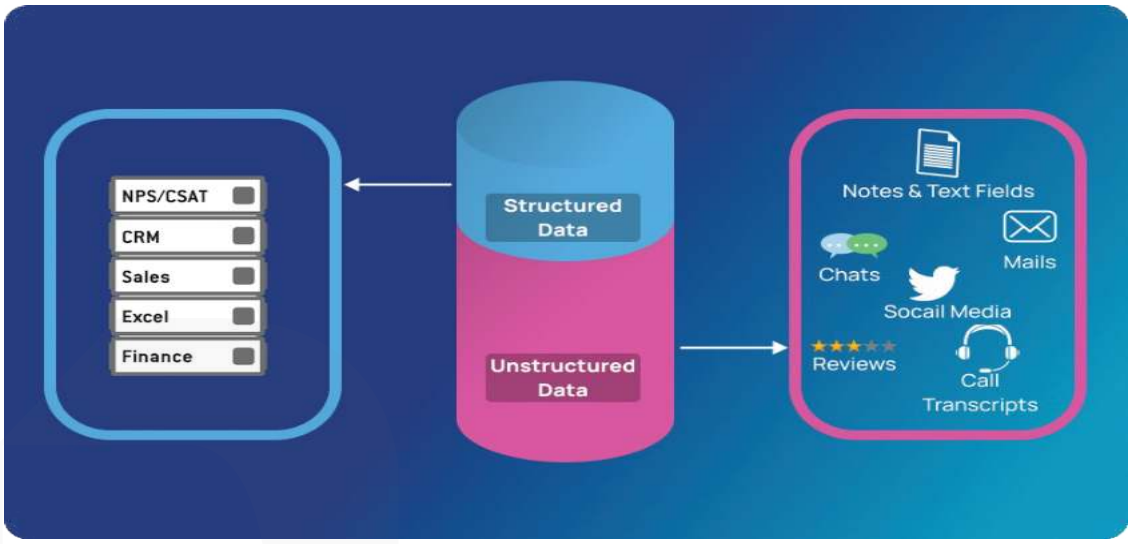
-تعقيد وغزارة التراكيب اللغوية

		الماضي		المضارع		الأمر
		المعلوم	المجهول	المعلوم	المجهول	
غائب مذكر	هو	كَتَبَ	كُتِبَ	يَكْتُبُ	يُكْتَبُ	
	هما	كَتَبَا	كُتِبَا	يَكْتُبَانِ	يُكْتَبَانِ	
	هم	كَتَبُوا	كُتِبُوا	يَكْتُبُونَ	يُكْتَبُونَ	
غائب مؤنث	هي	كَتَبَتْ	كُتِبَتْ	تَكْتُبُ	تُكْتَبُ	
	هما	كَتَبَتَا	كُتِبَتَا	تَكْتُبَانِ	تُكْتَبَانِ	
	هنّ	كَتَبْنَ	كُتِبْنَ	يَكْتُبْنَ	يُكْتَبْنَ	
مخاطب مذكر	أنتَ	كَتَبْتَ	كُتِبْتَ	تَكْتُبُ	تُكْتَبُ	اُكْتُبْ
	أنتما	كَتَبْتُمَا	كُتِبْتُمَا	تَكْتُبَانِ	تُكْتَبَانِ	اُكْتُبَا
	أنتم	كَتَبْتُمْ	كُتِبْتُمْ	تَكْتُبُونَ	تُكْتَبُونَ	اُكْتُبُوا
مخاطب مؤنث	أنتِ	كَتَبْتِ	كُتِبْتِ	تَكْتُبِينَ	تُكْتَبِينَ	اُكْتُبِي
	أنتما	كَتَبْتُمَا	كُتِبْتُمَا	تَكْتُبَانِ	تُكْتَبَانِ	اُكْتُبَا
	أنتنّ	كَتَبْتُنَّ	كُتِبْتُنَّ	تَكْتُبْنَ	تُكْتَبْنَ	اُكْتُبْنَ
متكلم	أنا	كَتَبْتُ	كُتِبْتُ	أَكْتُبُ	أُكْتَبُ	
	نحن	كَتَبْنَا	كُتِبْنَا	نَكْتُبُ	نُكْتَبُ	

تحديات معالجة اللغات الطبيعية للغة العربية

تحديات تقسيم الجمل
التعبيرات المجازية
الكلمات ذات الدلالات (القاهرة، العين)
الألفاظ المستحدثة
الغاية من الحديث (يا سلام)
المزاح

- كيفية معالجة اللغات الطبيعية :



خطوات معالجة اللغة الطبيعية (العربية)

"مدينة دمشق هي عاصمةُ الجمهورية العربية السورية، وأكبر المدن السورية من حيث الكثافة السكانية؛ تعد مدينة دمشق من أقدم العواصم المأهولة في العالم، وقد رجّح المؤرِّخون أن تاريخ المدينة يعود إلى ما قبل الألف السابع قبل الميلاد، حيث تم العثور على بعض الحفريات في منطقة تل الرماد، ودلّت هذه الحفريات أن تاريخ المدينة يعود إلى تسعة آلاف سنة قبل الميلاد. يعود اسم (دمشق) إلى أصولٍ آشوريّةٍ قديمةٍ، ويعني الأرض العامرة والزّاهِرة، دَلالةً على جمال طبيعتها وتضاريسها الخلّابة، ويُقال بأنها سُميت (شام) نسبةً إلى سام بن نوح عليه السلام. تقع مدينة دمشق في الجزء الجنوبي الغربي من الجمهورية السورية، يحدّ المدينة سهول حوران وجبال القلمون والبادية السورية، وتحيط بالمدينة بساتين الغوطة، وربوة دمشق، وجبل قاسيون، كما تطل المدينة على ضفاف نهر بردى".

- المرحلة الأولى: تجزئة الجمل (Sentence Segmentation)

- مدينة دمشق هي عاصمةُ الجمهورية العربية السورية، وأكبر المدن السورية من حيث الكثافة السكانية
- تعتبر مدينة دمشق من أقدم العواصم المأهولة في العالم.
- وقد رجَّح المؤرِّخون أن تاريخ المدينة يعود إلى ما قبل الألف السابع قبل الميلاد....

- المرحلة الثانية: الحصول على الوحدات اللغوية (Tokenization)

"مدينة"، "دمشق"، "هي"، "عاصمةُ"، "الجمهورية"، " العربية"، "السورية"، "و"، "أكبر"، "المدن"، "السورية"، "من"، "حيث"، "الكثافة"، "السكانية"، "."

تعتبر	دمشق	من	أقدم
"فعل"	"اسم علم"	"حرف جر"	"اسم"
العواصم	المأهولة	في	العالم
"اسم"	"اسم"	"حرف جر"	"اسم"

- **المرحلة الثالثة: التنبؤ بأقسام الكلام (Part of Speech)**

تعتبر	دمشق	من	أقدم
"اعتبر"	"دمشق"	"من"	"أقدم"
العواصم	المأهولة	في	العالم
"عاصمة"	"مأهول"	"في"	"عالم"

- **المرحلة الرابعة: أصل كلمات النص (Lematization)**

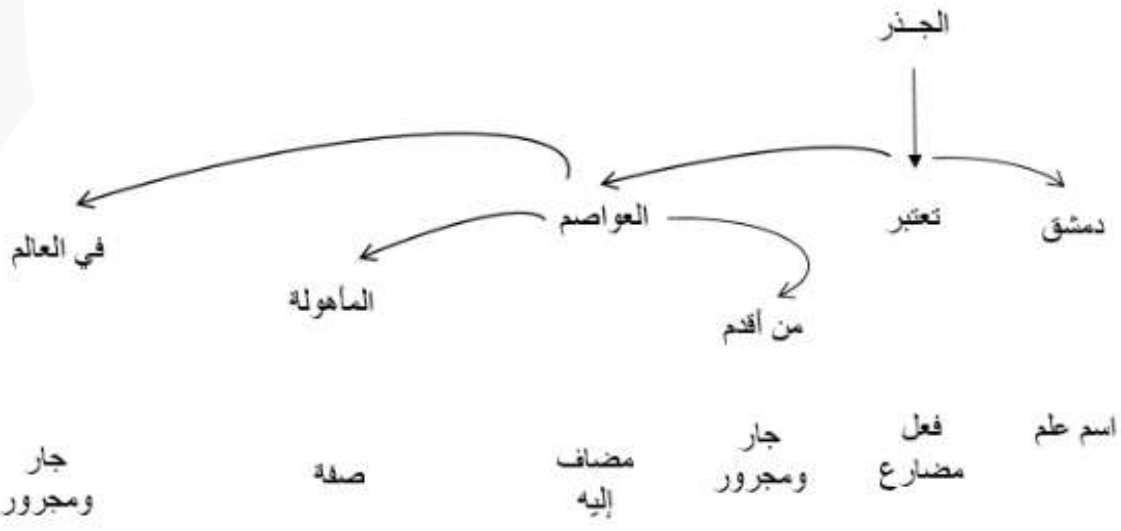
إرجاع كل كلمة إلى أصلها وجذرها.

تعتبر	دمشق	من	أقدم
"اعتبر"	"دمشق"	"من"	"أقدم"
العواصم	المأهولة	في	العالم
"عاصمة"	"مأهول"	"في"	"عالم"

- **المرحلة الخامسة: تحديد كلمات التوقف (Stop words removal)**

وهو نوع من المعالجة المسبقة للنص بإزالة وتصفية الكلمات الشائعة، أو الكلمات المستبعدة، وهي الكلمات التي تتكرر في النصوص مثل: (في ، من ، إلى ...)ويستحسن تجاهلها وعدم فهرستها.

- **المرحلة السادسة : الإعراب الاعتمادي (Dependencies parsing)**



- **المرحلة السابعة: التعرف على الكيانات المسماة (Named-Entity Recognition NER)**

دمشق	أقدم	العواصم	المأهولة
موقع جغرافي			
العالم	عاصمة	سورية	
		موقع جغرافي	

- **المرحلة الثامنـة: تحديد المرجعية للضمائر (Coreference Resolution)**

وتشتمل على عملية تحديد مرجعية ضمير معين ، وإشارته إلى اسم أو فعل أو فاعل معين، الرابط أدناه لأداة فراسة التي تعنى بتحديد المرجعية للضمائر

Farasa (qcri.org)

- **أنظمة ترميز اللغة العربية إلى الإنجليزية لمعالجتها ببرامج معالجة اللغات الطبيعية**

يتم ترميز حروف اللغة العربية باستخدام أنظمة ترميز معتمدة؛ لتحويل كل حرف إلى مقابل له باللغة الإنجليزية؛ لتسهل معالجته، واستخدامه بواسطة الحاسب

Buckwalter

SAMPA: Speech Assessment Methods Phonetic Alphabet

IPA : International Phonetic Alphabet

معالجة اللغات الطبيعية باستخدام حزمة ادوات (NLTK) مع لغة برمجة البايثون:

مجموعة أدوات اللغة الطبيعية (NLTK) هي عبارة عن منصة تستخدم لبناء برامج بايثون، التي تعمل مع بيانات اللغة البشرية للتطبيق في معالجة اللغة الطبيعية، حيث إن حزمة الأدوات تحتوي على مكتبات معالجة النصوص للترميز والتحليل والتصنيف والتجذير والتعليم المنطقي والدلالي. كما يتضمن أيضاً عروضاً توضيحيةً رسوميةً ومجموعات نماذج بيانات.

التحميل الصحيح للبايثون

تحميل آخر إصدار من بايثون من الموقع الرسمي

Download Python | Python.org

الدوال البرمجية بلغة بايثون لتطبيق تحويل النص إلى وحدات لغوية (Tokenization) :

- TreebankWordTokenizer
- WordPunctTokenizer
- PunctWordTokenizer
- WhitespaceTokenizer

الأوامر البرمجية بلغة بايثون لاستخراج وسوم أقسام الكلام (Part of speech tagging) من النصوص المدخلة:

GitHub - OmarQaisi/Part-Of-Speech-Tagger-for-Arabic-Language

المتطلبات :

تثبيت : Installing Pandas, matplotlib and xlrd

```
pip install pandas matplotlib xlrd<<
```

الأوامر البرمجية بلغة بايثون لاستخراج أصل الكلمات (تجذير)(Lemmatization) :

```
from nltk.stem.isri import ISRIStemmer
st = ISRIStemmer()
print st.stem(u'إعلاميون')
```

الأوامر البرمجية بلغة باثون لتحديد كلمات التوقف (Stop words removal)، وبعض المراجع لذلك

Arabic-Stopwords · PyPI

.GitHub - mohataher/arabic-stop-words: Largest list of Arabic stop words on Github

أكبر قائمة لمستبعدات الفهرسة العربية على جيت هاب

```
pip install Arabic-Stopwords<<
```

تطبيق الإعراب الاعتمادي (Dependencies parsing)

Dependencies parsing

تطبيق التعرف على الكيانات المسماة (Named-Entity Recognition NER)

Named-Entity Recognition NER

كيفية معالجة الرموز التعبيرية (Emoji) وتحويلها إلى نص

تتم معالجة الرموز التعبيرية في النصوص باستخدام الرمز الموحد المقابل للرمز (Unicode)

عن طريق استخدام المكتبات أو إنشائه.

```
message = input("> ")
words = message.split( " ")
emojis = {
   ":)" : "😀",
   ":(" : "😞",
   "lol" : "😂",
   "sick":"😰",
   "happy": "😃",
   "mermaid": "🧜‍♀️"
}
outcome = " "
for word in words:
   outcome += emojis.get(word, word) + " "
print(outcome)
```

معالجة اللغات الطبيعية وتعلم الآلة (NLP & Machine Learning)

تعتبر تقنيات تعلم الآلة (Machine Learning) من أهم التطبيقات وأكثرها ارتباطاً بمعالجة اللغات الطبيعية ويعد تعلم الآلة أحد فروع الذكاء الاصطناعي التي تهتم بتصميم وتطوير خوارزميات وتقنيات تسمح للحواسيب بامتلاك خاصية «التعلم».

تعتبر معالجة اللغات الطبيعة جزء فرعي من تعلم الآلة، ويركز على جعل الآلة قادرةً على تعلم لغة الإنسان، والتفاعل مع الكثير من التطبيقات التي تم شرحها مسبقاً، وتتطلب استخدام تقنيات تعلم الالة لبناء نموذج لغوي قادر على محاكاة لغة البشر. لمعرفة المزيد عن علاقة معالجة اللغات الطبيعية بتعلم الآلة، يمكن الاطلاع على الرابط التالي:

NLP & Machine Learning

أمثلة وتطبيقات على تحليل المشاعر (Sentiment Analysis example)

Sentiment Analysis example

أمثلة على قواعد البيانات الجاهزة (روابط)

قواعد بيانات جاهزة ومصممة لأغراض معالجة اللغات الطبيعية، يمكن الوصول لها على الرابط التالي:
(NLTK Data) موقع (كاجل) وهو أكبر موقع مخصص لقواعد البيانات الخاصة بالذكاء الاصطناعي، وتعلم الآلة، ويمكن الاستفادة منه لهذه الأغراض، ويمكن الوصول له عبر الرابط التالي:

Kaggle: Your Machine Learning and Data Science Community

في الختام :

- تعد تطبيقات معالجة اللغات الطبيعية من أهم التقنيات التي تضيف الكثير لأي مشروع أو منتج تقني؛ لأنها تسهل تفاعل التطبيق مع الإنسان، مما يجعل التطبيق أسهل وأكثر تفاعلاً مع الإنسان.

- تشكل تطبيقات معالجة اللغات الطبيعية تحدياً كبيراً في مجال الذكاء الاصطناعي؛ لأنها تعتمد على فك شفرة لغة الإنسان، وتراكيبها ومدلولاتها.

- تعتبر اللغة العربية من أكثر اللغات تحدياً في مجال الذكاء الاصطناعي ، ويعود ذلك إلى طبيعة اللغة العربية الفريدة المتنوعة في لهجاتها، واستخدام التشكيل، والتعقيد التركيبي للغة.

- إن اتباع خطوات المعالجة المسبقة للنصوص، ومعالجة اللغات الطبيعية من شأنه ان يجعل بناء تطبيقات معالجة اللغات الطبيعية أكثر سهولةً.

ما تمّ إنجازه في الكتيب:

- تمّ التعرف على مفاهيم وتقنيات معالجة اللغات الطبيعية، وتطبيقاتها المختلفة.
- أصبح لديك القدرة على تحديد، واختيار تطبيق معالجة اللغات الطبيعية المناسب لنوع المشروع ونشاط منشأتك.
- تم التعرف على أهم الأدوات، والبرامج المستخدمة في مجال معالجة اللغات الطبيعية.
- تم التعرف عن كثب على خطوات تطبيق معالجة اللغات الطبيعية للغة العربية، وكيفية تطبيقها على نص حقيقي.
- أصبح لديك قائمة بأهم المصادر، والمراجع التي تساعد على تطبيق تقنيات معالجة اللغات الطبيعية.

منشآت
monsha'at
الهيئة العامة للمنشآت الصغيرة والمتوسطة
Small & Medium Enterprises General Authority

مركز ذكاء

شكراً