Passengers

Natural languages

Powered by

Osama Al-Rashed
Mohannad kf alghazal

Final report

2022/6/21

.01

المحلل الصرفي

1. مقدمة Introduction

كان العرب قديماً يتحدثون اللغة العربيّة على سـليقتهم بطلاقة وفصـاحة دون الحاجة إلى قواعد تضـبط لغتهم، وخوفًا على اللغة العربية من الضياع قد جمعوها دون ترتيب ثم رُتبت على الألفاظ ومن ثمَّ توالت المعجمات بعد ذلك.

2. وصف عام Overview

يقوم هذا الملف بوصـــف محلل للّغة العربية، منها الصـــرفي والنحوي، حيث نعتمد بعملية التحليل على ما يســـمى بالمعجم ألا هو الكتاب الذي يحوي على شرحٍ المفردات والألفاظ اللغويّة وتوضيح معانيها وصفاتها ودلالاتها، واعتمادًا على نوع المحلل نقوم بتوليد الخرج المناسب للدخل المعطى.

الـ Data base الخاصة بالمشروع هي عبارة عن المعجم الذي يتألف من: ID, Word, Type

3. مواصفات المتطلبات الوظيفية

في هذا القسم سنتحدث عن المتطلبات الوظيفية للبرنامج، والجداول أدناه ستوضح هذا المتطلبات.

3.1 إدارة المعجم Manage Lexical [ML]

Requirement ID	الوصف
ML -1	يسمح النظام للمسؤول بـ "تصفح" المعجم.
ML -2	يسمح النظام للمسؤول بـ "بحث" عن كلمة بالمعجم.
ML -3	يسمح النظام للمسؤول بـ "تصفية" الكلمات.
ML -4	يسمح النظام للمسؤول بـ "إضافة" كلمة جديدة للمعجم.

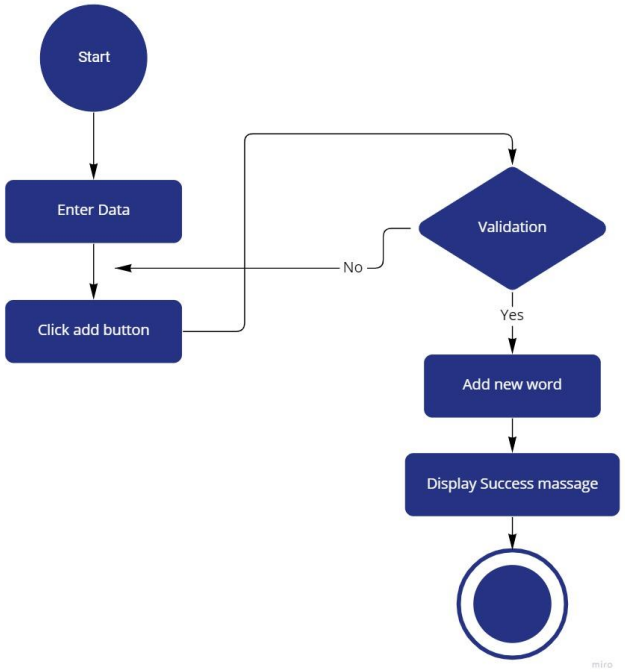
3.2 محلل الصرفي Morphological Analyzer [MA]

Requirement ID	الوصف
MA -1	يسمح النظام للمسؤول بـ "إيجاد" ناتج المحلل.
MA -2	يسمح النظام للمسؤول بـ "تهيئة" دخل وخرج المحلل.

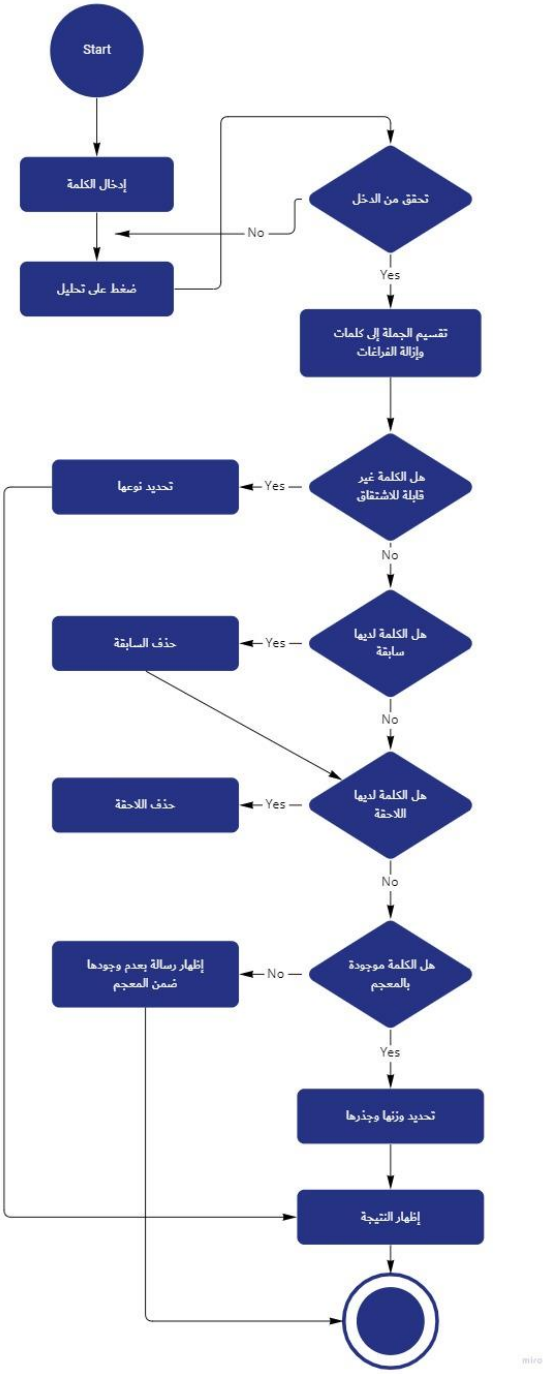
4. Activity diagram

تصف هذه المخططات آلية عمل الأفعال Actions داخل النظام.

4.1 إضافة كلمة جديدة إلى المعجم [ML-4]



4.2 إيجاد ناتج لمحلل الصرفي [MA-1]



.02

المحلل الصوتي

1. مقدمة Introduction

يمثل الإحساس المتولد عن كل حاسة من الحواس الخمسة بالنمط (Mode) خاصًا للإدراك، فالإدراك الناتج عن السماع شيء هو النمط السمعي، الذي يقوم بتمييز الأصوات المتباينة، وذلك بوجود وسط ليتم نقل الصوت إلى الآذان لكي تتم عملية السمع عن طريق نقل الإهتزازات إليها.

حيث يكون هذا الكلام إما مكتوبًا مقروءًا أو منطوقًا مسموعًا، ودراسة الكلام المنطوق المسموع لابد منه لدراسة الأنظمة اللغوية، ويعتمد الكلام المنطوق المسموع على أساسين: حركي (مخارج) وسمعي (صفات) ويؤدي تمييز الأصوات إلى بناء نظام صوتي لغوي، لكل حرف صوتٌ خاص وللتعبير عن كلمة ما يتم لفظ الأصوات المرتبطة بالحروف لفظًا متتابعًا بنفس الترتيب الكتابي للكلمة، وبهذا يتم الحصول على الكلمة صوتيًا.

2. وصف عام Overview

يقوم هذا الملف بوصف محللًا صوتيًا للمحارف العربية، حيث نعتمد بعملية التحليل على صفات الصوت المنطوق (RMS, ZCR, Energy) ونطابق تلك الصفات على الأصوات المخزنة مسبقًا في قاعدة الأصوات (المسموعة المنطوقة، والمكتوبة المقروءة) ومن ثمَّ يتم التعرف على هذا الصوت.

دعونا نتعرف على هذه الصفات Parameters لنتمكن من فهم كيفية تحليل الصوت.

2.1 Root Mean Square [RMS]

تحسب هذه القيمة لمجموعة من العينات، وهو القيمة المتوسطة للجذر التربيعي لمجموعة القيم (العينات) لشكل التي تتوافق مع جهارة الصوت، وتعطى بالعلاقة

$$\frac{1}{N}\sqrt{\sum_{i=1}^{n}(x)^2}$$

2.2 Zero Crossings [ZCR]

يقصد بها عدد المرات التي تعبر فيها الإشارة من قيم سالبة لموجبة، والعكس، أما معدلها فهو عدد المرات خلال ثانية واحدة.

2.3 Energy

هو مقدار قوة الاهتزازات الناتجة عن التسجيل الصوتي.

3. مواصفات المتطلبات الوظيفية

في هذا القسم سنتحدث عن المتطلبات الوظيفية للبرنامج، والجداول أدناه ستوضح هذا المتطلبات.

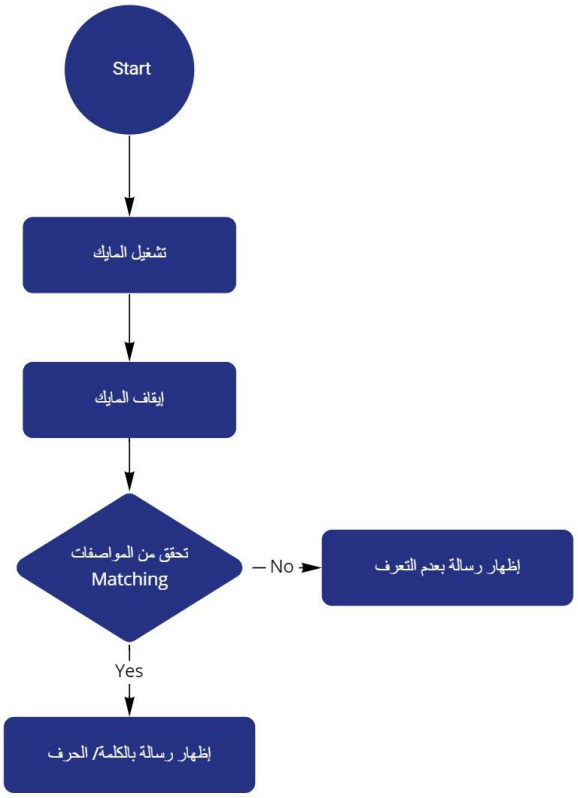
3.1 إدارة الأصوات Manage Audio [MA]

Requirement ID	الوصف
MA -1	يسمح النظام للمسؤول بـ "استعراض"الأصوات.
MA -2	يسمح النظام للمسؤول بـ "إضافة" صوت جديد.
MA -3	يسمح النظام للمسؤول بـ "مراجعة" الصوت.
MA -4	يسمح النظام للمسؤول بـ "فلترة" الأصوات.
MA -5	يسمح النظام للمسؤول بـ "التعرف" على صوت.

4. Activity diagram

تصف هذه المخططات آلية عمل الأفعال Actions داخل النظام.

4.1 التعرف على الصوت [MA-5]



.03

المركب الصوتي

1. مقدمة Introduction

تتكون اللغة في أساسها من الأصوات التي تسمى حروفًا؛ فالأصوات قد يؤثر بعضها على بعض حين تتجاور داخل الكلام لتتراكب الأصوات مع بعضها، بالإضافة إلى أنَّ الكلام أداء رمزي في إطار اجتماعي ما، وهذا الإطار الاجتماعي هو اللغة.

2. وصف عام Overview

للتركيب الصوتي منحيان، المنحى الأول هو تركيب الإشارات الصوتية مع الاعتماد على المنظومة اللغوية، حيث يتم توليد الكلام المستمر من غير التقيد بعدد المفردات، أما المنحى الثاني هو تركيب الإشارات الصوتية فيزيائيًا من دون الاعتماد على المنظومة اللغوية، ويتم ذلك عن طريق المحلل الصوتي.

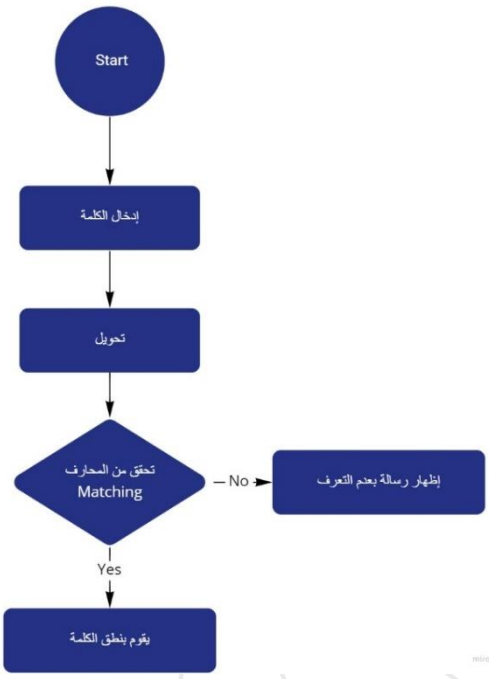
3. مواصفات المتطلبات الوظيفية

في هذا القسم سنتحدث عن المتطلبات الوظيفية للبرنامج، والجداول أدناه ستوضح هذا المتطلبات.

3.1 إدارة الأصوات Manage Audio [MA]

Requirement ID	الوصف
MA -1	يسمح النظام للمسؤول بـ "استعراض"الأصوات.
MA -2	يسمح النظام للمسؤول بـ "إضافة" حرف جديد.
MA -3	يسمح النظام للمسؤول بـ "تحويل" الكلمة إلى صوت.

4. Activity diagram



.04

المحلل النحوي

1. مقدمة Introduction

إنَّ بناء محلل نحوي للّغة عمل متعدد الجوانب، يصـــب فيه نتاج كثير من النظريات النحوية الحديثة، وأســـاليب الذكاء الإصـطناعي المتطورة، حيث تمثل عملية التحليل النحوي النحوي إحدى الموقومات الرئيسـية لنطق النصـوص آليًا (TTS) وهي إحدى التطبيقات التي تتعاظم أهميتها مع التوسع في استخدام أساليب الذكاء الإصطناعي والأنظمة الأخرى. ويقوم المحلل النحوي بإيجاد القالب النحوي للجملة وفقًا لقواعد اللغة العربية.

يمكن تحديد مكونات النظام النحوي بالنقاط التالية:

1- طائفة من المعاني النحوية العامة (الخبر، الإنشاء، النفي...).
2- طائفة من المعاني النحوية الخاصة (الحال، الفاعل، المفعول به...).
3- مجموعة من العلاقات التي تربط بين المعاني الخاصة كعلاقة الإسناد والتخصيص والنسبة والتبعية.
4- المباني الطالحة التي قدمها علما الصرف والصوتيات لعلم النحو.
5- القيم الأخلاقية "المقابلات" (الخبر مقابل الإنشاؤ، المدح مقابل الذم).

05.

الويب الدلالي

1. مقدمة

Semantic web عبارة عن شــبكة من البيانات المرتبطة بطريقة معينة يمكن من خلالها معالجتها بســهولة بواســطة الآلات، بدلًا من العامل البشــري، بالإضــافة إلى أنها وســيلة فعالة لتمثيل قاعدة البيانات والمعلومات.

يهدف Semantic web لتحويل البيانات المهيكلة (Trees) إلى بيانات غير مهيكلة، بالإضــافة إلى تمكين المستخدمين من البحث عن معلومات واكتشافها ومشاركتها أيضًا.

هناك العديد من المهام التي يقوم بتنفيذها البشــر من خلال صــفحات الويب لكن من الصــعب الاعتماد على الآلة بشكل كامل لأن صفحات الويب مصممة ليقرأها البشر وليس الآلة.

لدى Sematic web رؤية مستقبلية حيال تفسير البيانات وتحليلها حيث يعد هذا الأمر شاقًا بالنسبة لبني البشر مقارنةً بالســرعة التي تتصــف بها الآلة بتفســير البيانات مما يســمح لها بتنفيذ العديد من المهام المتعلقة باكتشاف المعلومات ومزجها واتخاذ قرار أو إجراء حيالها.

من تطبيقات Semantic web، هو Semantic search ألا وهو البحث الدلالي فهو أســلوب للبحث عن البيانات، يعتمد في عمله على السياق والمضمون والهدف ومفهوم العبارة التي يتم البحث عنها، ويشتمل البحث الدلالي أيضًــا على الموقع ومرادفات المصــطلح والاتجاهات الحالية واختلافات الكلمات وعناصــر اللغة الطبيعية الأخرى كجزءٍ من البحث.

هذا هو الســبب الذي يجعلك عندما تكتب "مطاعم" في محرك البحث الخاص بك ، فإنها تمنحك قائمة بالمطاعم القريبة.

بطريقة ما، هو يســهيل الانتقال بين الطريقة التي يتفاعل بها المســتخدمون مع الأشــخاص مقابل الطريقة التي يتفاعلون بها مع نتائج البحث.

لذلك، يضيف البحث الدلالي مستوى من الفهم للاستعلامات، حيث يتم اشتقاق مفاهيم البحث الدلالي من خوارزميات ومنهجيات بحث مختلفة، بما فيها تعيين كلمات مفتاحية، Graph patterns، Fuzzy logic، ولكن هذه الخوارزميات لها أيضًا أنماط تعلم.

من خلال معدلات الارتداد ومعدلات التحويل وأنواع أخرى من المؤشرات، يمكن لهذه الخوارزميات تحسين رضا المستخدم لمطابقة الكلمات الرئيسية والصفحات بشكل أفضل.

لذلك، يرتبط البحث الدلالي ارتباطًا وثيقًا بالتعلم الآلي، من حيث أنه يســتخدم البيانات الســابقة وأنماط التجربة والخطأ لتحسين تجربة المستخدم.

بالإضافة إلى أنَّ الخوارزميات تعلم الآلة والتعلم العميق لا تستطيع التعامل مع البيانات النصية الخام، لذلك لا بد من تحويل هذه البيانات النصـية إلى أشـعة رقمية، وتعد تقنيات Vectorization من العمليات الأسـاسـية في علم المعطيات، في هذا النموذج كل كلمة في النص تقابل رقمًا صحيحًا.

يجب تطبيق عملية تقطيع النص Tokenization قبل تطبيق عملية Vectorization، وإجراء بعض العمليات الإختيارية كالتجزئة Segmentation والتجذيع Stemming وإيجاد فروع للكلمات Lemmatization.

2. Text Vectorization Methods

2.1. حقيبة الكلمات Bag of words

في هذه الخوارزمية يتم الإشـــارة إلى الكلمة بفهرس Index ويحوّل كل كلمة إلى عدد تكرراتها، هذه الخوارزمية لا توضح تَوضُّع الكلمة، بل توضح التكررات فقط، كل نص يُوضَح بشعاع من ترددات، ويطلع على هذا التمثيل بـــــ Term frequency vector، وتعتبر هذه الخورزمية مكلفة نظرًا لعدد مرات طلب الولوج إلى المعطيات.

2.2. تردد الحد – تردد المستند العكسي TF-IDF

تم طرح هذه الطريقة لحل مشــكلة التكرار عن طريق تحديد مقياس أهمية الكلمة، حيث يأخذ بعين الاعتبار طول النص وتكرار الكلمة في المجموعة النصية Corpus.

بداية دعونا نقوم بحساب TF وهو معيار يعبر عن تكرار الكلمة في فجموعة النصوص.

$$tf\left(\text{الكلمة}\right) = \frac{\text{النص في الكلمة تكرار مرات عدد}}{\text{النص لكلمات الكلي العدد}}$$

أما معيار IDF فيحسـب بتقسـيم عدد النصـوص الكلية في المجموعة على عدد النصـوص التي تحتوي الكلمة ثم يدخل الناتج على تابع لوغاريتم.

$$IDF\left(\text{الكلمة}\right) = \log\frac{\text{للنصوص الكلي العدد}}{\text{الكلمة تحتوي التي النصوص عدد}}$$

بالنسبة للكلمات والأحرف الشائعة تكون قيمة TF كبيرة، أما قيمة IDF تكون صغيرة ومتقاربة للصفر، للحصول على TF-IDF من العلاقة:

$$TF\_IDF\ \left(\text{الكلمة}\right) = TF\ \left(\text{الكلمة}\right) * IDF\ \left(\text{الكلمة}\right)$$

كل نص يمثل بشعاع بعده يساوي عدد كلمات الـ Vocabulary المستخرج من المجموعة النصية. تمتاز هذه الطريقة بكفاءة أعلى من Bag of words.

2.3. كمية المحارف N-Grams

تعتمد هذه التقنية على احتمال وجود مجموعة من الكلمات عددها n مع بعضها في مجموعة النصوص، وتعتمد هذه الطريقة على أيضًا على السياق؛ لأنه وفقًا للغة العربية فهناك كلمات تستخدم ضمن سياق واحد وأخرى فلا.

تمتاز هذه الطريقة عن قرينتها TF-IDF حيث أنَّ الأخيرة تفقد المعلومات الموجودة في جوار الكلمات، في حين أنَّ N-Grams تمتاز بالتنبؤ من خلال الكلمات المجاورة.

2.4. تخطي كمية محارف Skip-gram

هي طريقة جزئية من N-Grams لكن تعتمد بشكل رئيسي على قيمة الخطوة Skip، وقيمة n من تحديد حجم المجموعات.

2.5. تشفير واحد جديد One hot – encoding

طريقة بسيطة لتمثيل الكلمات اعتمادًا على شعاع الكلمات، يتم إنشاء vocabulary يحتوي على الكلمات الأكثر شيوعًا في مجموعة النصوص، حيث يمثل كل منها بشعاع طوله مساويًا لعدد كلمات vocabulary، تكون عناصر الشعاع مساوية للصفر ماعدا القيمة المقابلة لمكان الكلمة في الـــ vocabulary فتكون قيمتها واحد.

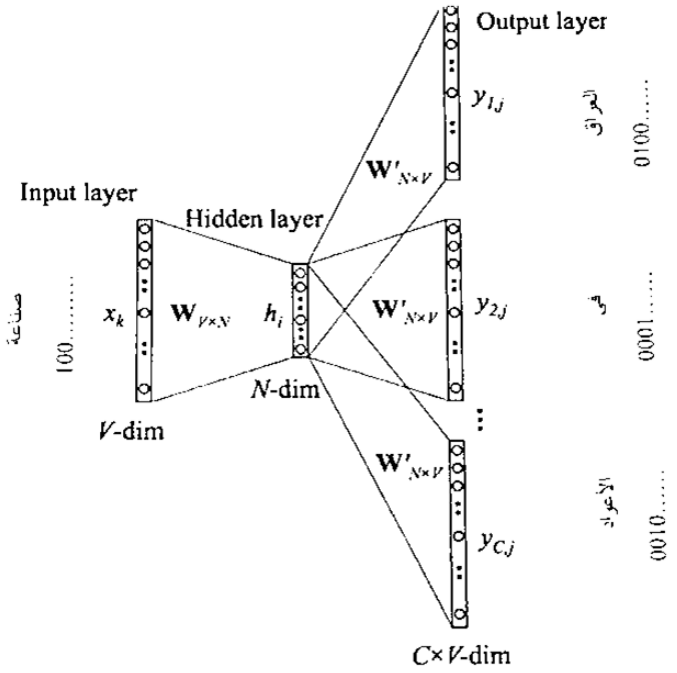
2.6. تضمين الكلمات Word Embedding

هذه الطريقة جاءت لتحل مشكلة الطريقة السابقة one-hot التي تشفر على مستوى كلمة واحدة فقط بالتالي إذا كان النص طويل تصبح شيفرة الكلمة الواحدة طويلة جدًا.

تقدم هذه طريقة تمثيلًا كثيفًا Dense للكلمات ومعانيها نسبة إلى السياق الذي تأتي به عادة. إنّ هذه الطريقة تضغط أشعة one-hot الطويلة إلى أشعة أصغر طولًا وتحمل معنى السياق.

2.6.1. Word-2-vec-Skipgram

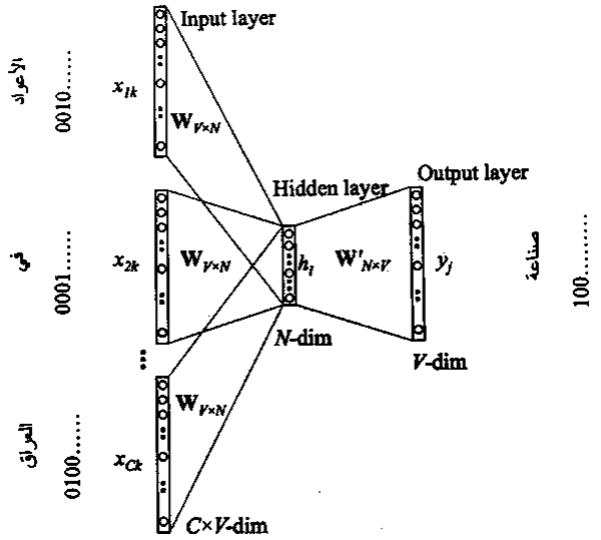
نقوم بتدريب شبكة عصبية على نص ما ليكن: صناعة الأعواد في العراق.



بعد التدريب أوزان الشبكة تكون شيفرة المطلوب لكلمة "صناعة" لكن ليس جميع الأوزان وإنما فقط الأوزان الموصولة مع العصبون ذات قيمة 1 في الشيفرة one-hot-encoding.

Word-2-vec continues bag of work .2.6.2

تختلف عن سابقتها بشكل الشبكة العصبية فقط كم هو موضح.



3. مراحل ال Semantic web

Named Entities هو الكيان الموصوف في المستند، يساعد الكيان الحاسوب على فهم كل ما تعرفه عن شخص أو مؤسسة أو مكان مذكور في مستند، ومن أمثلته أسماء العلم.

3.1. **التعرف على الكيلنات المسماة (NER)** هو تطبيق لمعالجة اللغة الطبيعية (NLP) للذي يعالج ويفهم كميات كبيرة من اللغة البشرية غير المنظمة، يُعرف أيضًا باسم تعريف الكيان وتقسيم الكيان واستخراج الكيان، استخلاص NER هو الخطوة الأولى في الإجلبة على الأسئلة واسترجاع المعلومات ونمذجة الموضوع، وهناك العديد من النماذج لتطبيق NER حسب حاجة التطبيق.

3.2. **تصنيف الكيانات المسماة (NEC)** هي عملية تصنيف الكائنات إلى فئات محددة، الهدف هو تطوير تقنيات عملية ومستقلة عن المجال من أجل اكتشاف الكيانات المسماة بدقة عالية تلقائيًا.

3.3. **ربط الكيان**، يُشار إليه أيضًا بربط الكيان المُسمى توضيح الكيان المُسمى (NED)، هي مهمة تعيين هوية فريدة للكيلنات، بالنظر إلى الآية القرآنية "وما أظن الساعة قائمة"، فإن الفكرة هي تحديد أن "الساعة" تشير إلى يوم القيامة وليس مقدار ساعة من الزمن.

نرى أنَّ طرق استخراج المعلومات (IE) مثل الكيان المسمى الاعتراف (NER)، تصنيف الكيان المسمى (NEC)، اسمه ربط الكيان، واستخراج العلاقة (RE)، والاستخراج الزمني، يمكن أن يساعد استخراج الحدث في إضافة ترميز إلى صفحات الويب.

4. علاقة الويب الدلالي بالنتولوجي

تعتبر اللبنة الأساسية للويب الدلالي لأنها تسمح بتفسير البيانات المدعومة آليًا مما يقلل من مشاركة الإنسان في البيانات وتكامل العمليات.

الأنطولوجيا "هي تحديد رسمي وصريح لمفهوم مشترك. يشير إلى نموذج مجرد للظواهر في العالم من خلال تحديد المفاهيم ذات الصلة لتلك الظواهر. يعني الصريح أن نوع المفاهيم المستخدمة والقيود المفروضة على استخدامها محددة بوضوح رسمي يشير إلى حقيقة أن الأنطولوجيا يجب أن تكون قابلة للقراءة آليًا المشتركة تعكس تلك الأنطولوجيا.

عندما يتم تمثيل المعرفة حول مجال ما بلغة تعريفية، فإن مجموعة الأشياء التي يمكن تمثيلها تسمى عالم الخطاب، يمكننا كتابة علم الوجود لبرنامج ما من خلال تحديد مجموعة من المصطلحات التمثيلية. تربط التعريفات أسماء الكيانات في عالم الخطاب (مثل الفئات أو العلاقات أو الوظائف أو الأشياء الأخرى) بالنص الذي يمكن للبشر قراءته والذي يصف ماهية الأسماء.

إنَّ مجموعة من خدمات الويب التي تشترك في نفس الأنطولوجيا ستكون قادرة على التواصل حول مجال الخطاب.