الإحصاء والبيانات اللغوية

https://www.youtube.com/watch?v=2rt2nTyFeDg&t=5s

د. سلطان المجيول
@Arabic_CL

الإحصاء والبيانات اللغوية (لغة آر)

أولا: ما المقصود بالبيانات اللغوية؟

ثانيا: تعلم الآلة للبيانات اللغوية (اختيار النموذج الأفضل للبيانات اللغوية) [https://machinelearningmastery.com/]

ثالثا: الإحصاء (ما المشكلات؟ ما الافتراضات؟ ما المشاريع التي يُمكن أن تعالج وتحلل باستعمال لغة البرمجة آر)

- التكرار والتشتت والشيوع
- علم الدلالة والمعاني
- النحو المعجمي: الانحدار اللوجستي
- البيانات اللغوية التاريخية: تحليل التقلب الاستعمال والقمم والقعور
- التنوع اللغوي: التجميع
- التنوع اللغوي: الارتباط

رابعا: ملفات التدريب لإعداد مشاريع في معالجة البيانات (اللغوية)

الآر أم البايثون؟

	الآر	البايثون
فئة المستعملين	المبرمجون/ علماء البيانات/ المحللون للبيانات/ الأكاديميون والباحثون/ المطورون	المبرمجون/ علماء البيانات/ المحللون للبيانات/ الأكاديميون والباحثون/ المطورون
المكتبات	dplyr, tidyr, data.table (manipulate data) stringr (manipulate strings datasets, mlbench, e1071, etc. (explore and perfume statistical tests) ggplot2 (visualization) carete (machine learning)	pandas (manipulate data) SciPy and NumPy (scientific computing) statsmodels (explore data, estimate models, perform statistical tests) Matplotlib and seaborn (graphics) Scikit-learn (machine learning)
حالات الاستخدام (تحليل البيانات)	مناسب للإحصائيين، وأكثر متعة وفعالية في تمثيل البيانات الدالة الواحدة قد تُكتب بأكثر من طريقة (الأكواد معقدة)	مناسب لمهندسي البيانات، ومفيد لما هو أكثر من علم البيانات وتحليلها الأكواد سهلة التعلم

المصدر: https://www.datacamp.com/community/tutorials/r-or-python

مبادئ أساسية لعمل مشاريع للبيانات اللغوية

- إتقان علم الإحصاء يمنحك مزيدًا من التمكن.
- قد يجعلك الإحصاء مثبطا لأن الأدوات والتحليلات الإحصائية قد أصبحت أكثر تشابكًا وتعقيدًا؛ لأن التحليل الإحصائي يتضمن العديد من المسارات. نحتاج إلى اختيار بيانات لغوية مناسبة وتقنية تحليلية فعالة وتفسير مناسب للنتائج.
- لا تحتاج إلى معرفة الخوارزميات الرياضية فالحاسب هو الملاذ لاستخلاص المعلومات من البيانات اللغوية.
- لا تكن ضحية للتقنيات الإحصائية التجارية في السوق الإحصائية.
- تذكر دائما بأن الأسلوب الإحصائي الأقوى هو المنطق السليم.
- النتائج تعتمد على ما تهيئه من بيانات لغوية.
- ركز على مشروع واحد.
- **المشاريع في تحليل البيانات اللغوية (العربية) عديدة. النصيحة المهمة: لا تترك مشروعا لم تُنهه وتدخل في مشروع آخر.**

Jason Brownlee

Machine learning is taught by academics, for academics.
That's why most material is so *dry* and *math-heavy*.

Developers need to know what works and how to use it.
We need *less math* and *more tutorials with working code*.

Me: Let's have fun…

Fun makes us more optimistic in my progress

ماذا نعني بالبيانات اللغوية؟

- تُحول البيانات اللغوية النصية الخام المكتوبة أو المنطوقة إلى بيانات تتضمن string، وnumeric، وinteger، decimal، في رأس الجدول

- تتضمن instances أو observations (صفوف) وattributive أو variables (أعمدة) في ملفات مفصولة متغيراتها وقيمها في جداول:
  csv، وrelational database tables ، وJSON، وfixed-width formatted file، وcalc…

- يتضمن الـ strings حروفا أو أرقاما أو ترميزات، أما الـ numeric فيتضمن أرقاما فقط (البيانات اللغوية)، أو integer.

- المتغير variable: يكون حدثا أو فكرة أو مرحلة زمنية يراد قياسها
- المتغير اللغوي: أسماء أو صفات أو ضمائر أو عبارات إلخ.

(كلما زاد الوقت في التعلم [مستقل] زادت الدرجات في الاختبار [تابع]): الوقت يغير الدرجة والعكس غير صحيح

1. المستقل: لا يتأثر بالتابع الخاضع للقياس: الجنس، أو العمر، أو الوعاء غير المتأثر بالتابع.
2. التابع: يتأثر؛ مثلا: التعريف والتنكير بين الذكر والأنثى، أو الصفات والأفعال والأسماء والضمائر وأدوات الربط بين الأوعية والمجالات والموضوعات.

ماذا نعني بالبيانات اللغوية؟

- أنواع المتغيرات

– المتغيرات اللغوية مقابل المتغيرات المستقلة (التأويلية):

1. **المتغير الاسمي nominal/ الفئوي categorical**

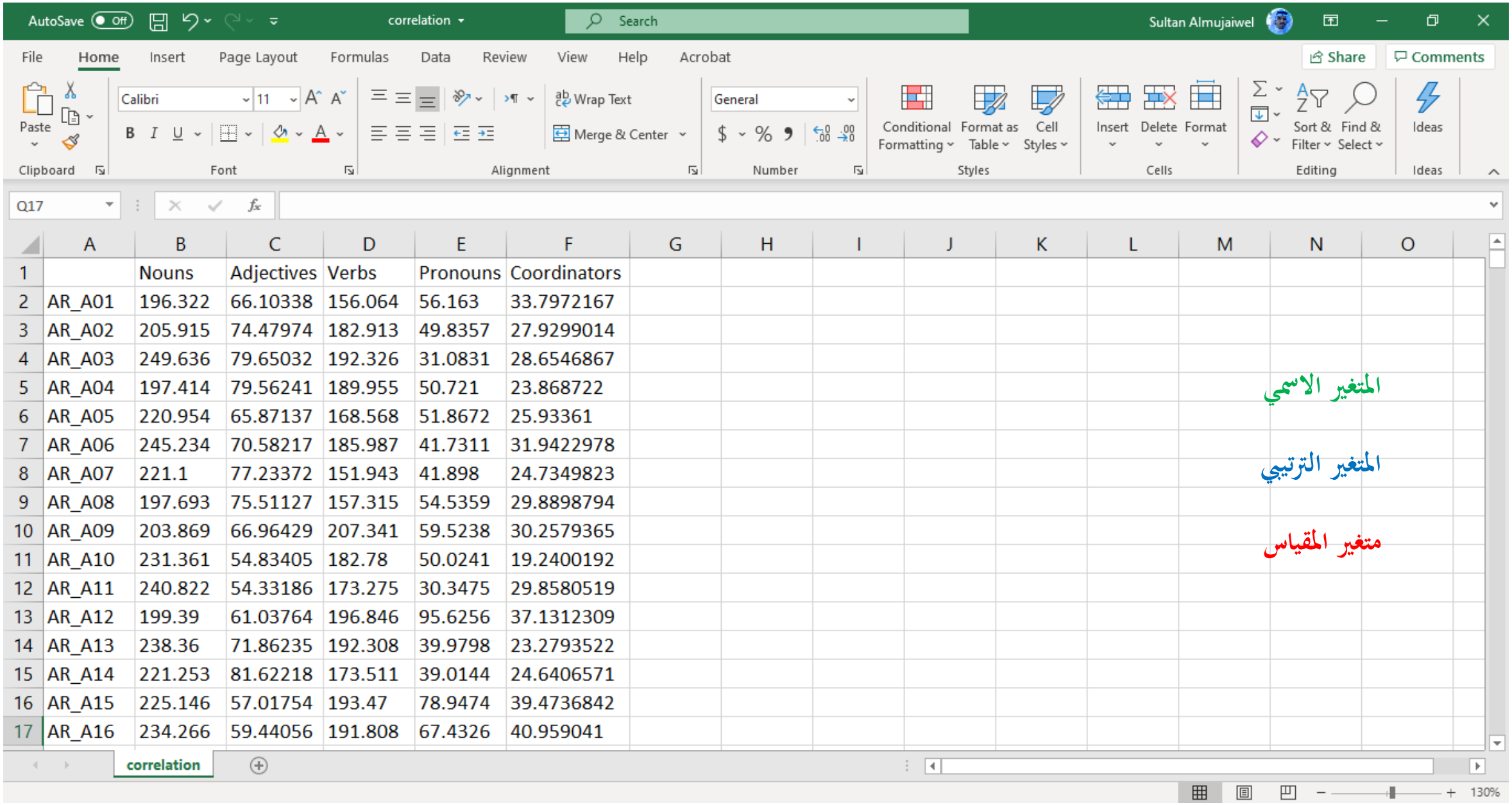
2. **المتغير الترتيبي ordinal/ الرتب ranks**

3. **متغير المقياس scale (ratio)**

- هل هناك علاقة بين جنسي المتحدثين (متغير تأويلي اسمي) واستخدام الضمائر الشخصية (متغير لغوي مقياسي)؟
- هل لكفاءة المتحدث اللغوية (متغير ترتيبي) تأثير على استخدام الكلمات الطويلة (متغير لغوي مقياسي)؟
- هل بين استخدام ضمير المتكلم واستخدام ضمير المخاطَب (وكلاهما متغيران لغويان مقياسيان) علاقة؟

4. المتغير الناتج (المتغير اللغوي مثل: الخصائص النحوية) ومتغير المتنبئ (المتغير المستقل مثل: المتغيرات السياقية المرتبطة بتلك الخصائص)

ماذا نعني بالبيانات اللغوية؟ (بنية البيانات)



	A	B	C	D	E	F
1		Nouns	Adjectives	Verbs	Pronouns	Coordinators
2	AR_A01	196.322	66.10338	156.064	56.163	33.7972167
3	AR_A02	205.915	74.47974	182.913	49.8357	27.9299014
4	AR_A03	249.636	79.65032	192.326	31.0831	28.6546867
5	AR_A04	197.414	79.56241	189.955	50.721	23.868722
6	AR_A05	220.954	65.87137	168.568	51.8672	25.93361
7	AR_A06	245.234	70.58217	185.987	41.7311	31.9422978
8	AR_A07	221.1	77.23372	151.943	41.898	24.7349823
9	AR_A08	197.693	75.51127	157.315	54.5359	29.8898794
10	AR_A09	203.869	66.96429	207.341	59.5238	30.2579365
11	AR_A10	231.361	54.83405	182.78	50.0241	19.2400192
12	AR_A11	240.822	54.33186	173.275	30.3475	29.8580519
13	AR_A12	199.39	61.03764	196.846	95.6256	37.1312309
14	AR_A13	238.36	71.86235	192.308	39.9798	23.2793522
15	AR_A14	221.253	81.62218	173.511	39.0144	24.6406571
16	AR_A15	225.146	57.01754	193.47	78.9474	39.4736842
17	AR_A16	234.266	59.44056	191.808	67.4326	40.959041

المتغير الاسمي

المتغير الترتيبي

متغير المقياس

خوارزميات تعلم الآلة

- **تعلم الآلة غير الموجه unsupervised والموجه supervised:**

1. **غير الموجه: التجميع.**
2. الموجه: تصنيف classification والمخرج منه عبارة عن متغيرات فئوية، أو انحدار regression (الخطي وغير الخطي) والمخرج منه عبارة عن متغيرات رقمية
3. **نماذج دقة التصنيف**

- ✓ Linear Discriminant Analysis LDA (simple linear)
- ✓ Classification and Regression Trees (CART) <- nonlinear
- ✓ k-Nearest Neighbors (KNN) <- nonlinear
- ✓ Support Vector Machine (SVM) (nonlinear)
- ✓ Random Forest (RF) <-complex nonlinear
- Regularized Regression
- Naïve Bayes
- Decision Trees <-complex nonlinear
- Linear Regression
- Logistic Regression (mixed effects models)

اختيار النموذج الأفضل في تعلم الآلة (التصنيف)

- عند البدء في مشروعك المحدد، قد تتساءل عن بعض النماذج، وأيها يصلح؟

- هناك أربعة أساليب لتقدير وتقييم أداء النماذج

الأول: تقسيم البيانات إلى بيانات تدريب 80% وبيانات اختبار 20%

الثاني: استخدام أسلوب التمهيد bootstrap

✓ الثالث: استعمال طريقة التحقق من الدقة بتقسيم البيانات إلى عدة أجزاء k-fold cross-validation

الرابع: استخدام طريقة leave one out cross-validation [التحقق من الدقة باختبار كل instance (صف) مع بقية الصفوف باعتبارها مجموعة بيانات تدريبية .

اختيار النموذج الأفضل في تعلم الآلة (التصنيف)

- من الطرق التي تساعد على تقييم الخوارزمية بدوالها ومكتبتها في الآر ما يعرف بـ k-fold cross-validation. يشير الـ k إلى عدد مرات تقويم البيانات، وإعطاء فكرة أولية عن دقة المتنبئات
- تقيم نتائج النموذج بالتحقق من دقة المتوسط، وذلك بتقييم كل نموذج 10 مرات [10-fold] بين بيانات التدريب والاختبار 80% وبيانات التحقق 20%

```
# Run algorithms using 10-fold cross-validation
trainControl <- trainControl(method="cv", number=10)
metric <- "Accuracy“

# LDA
set.seed(7)
fit.lda <- train(Species~., data=dataset, method="lda", metric=metric,
trControl=trainControl)

# CART
set.seed(7)
fit.cart <- train(Species~., data=dataset, method="rpart", metric=metric,
trControl=trainControl)

# KNN
set.seed(7)
fit.knn <- train(Species~., data=dataset, method="knn", metric=metric,
trControl=trainControl)

# SVM
set.seed(7)
fit.svm <- train(Species~., data=dataset, method="svmRadial", metric=metric,
trControl=trainControl)

# Random Forest
set.seed(7)
fit.rf <- train(Species~., data=dataset, method="rf", metric=metric, trControl=trainControl)
```

اختيار النموذج الأفضل في تعلم الآلة (أر أنموذجا تطبيقيا)

- استخلاص دقة كل نموذج

```
Models: lda, cart, knn, svm, rf
Number of resamples: 10
```

Accuracy

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
lda	0.5833333	0.6666667	0.6666667	0.6833333	0.7291667	0.8333333	0
cart	0.5833333	0.7500000	0.8333333	0.8000000	0.8958333	0.9166667	0
knn	0.5000000	0.6666667	0.7500000	0.7083333	0.7500000	0.9166667	0
svm	0.5833333	0.6666667	0.7500000	0.7416667	0.8333333	0.8333333	0
rf	0.6666667	0.7708333	0.8333333	0.8166667	0.8333333	0.9166667	0

Kappa

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
lda	0.375	0.50000	0.500	0.5250	0.59375	0.750	0
cart	0.375	0.62500	0.750	0.7000	0.84375	0.875	0
knn	0.250	0.50000	0.625	0.5625	0.62500	0.875	0
svm	0.375	0.50000	0.625	0.6125	0.75000	0.750	0
rf	0.500	0.65625	0.750	0.7250	0.75000	0.875	0

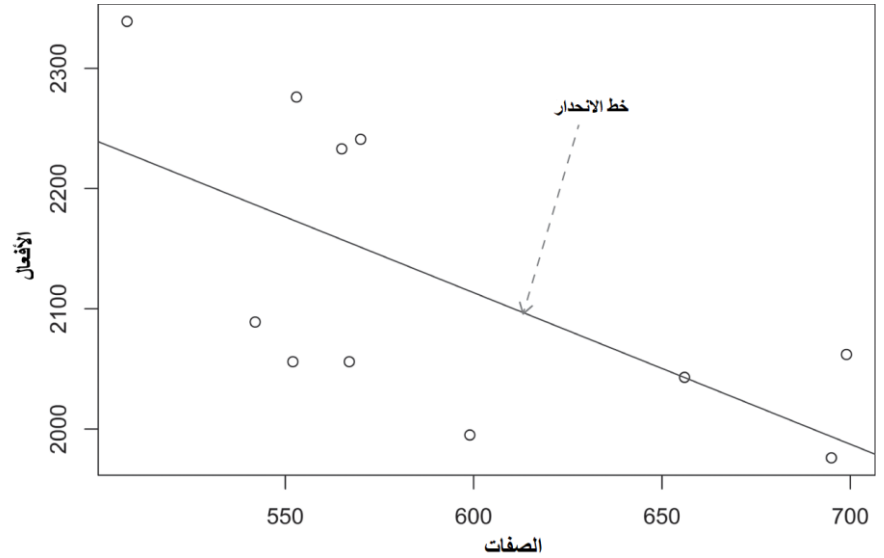
- الغابة العشوائية rf نموذج فعال للقيم المستمرة المقيسة measured
- تحليل المحدد الخطي lda يصلح للمتغيرات المستقلة المستمرة والمتغيرات التابعة التصنيفية class أو categories (عكس ANOVA). كذلك يصلح للقيم المنفصلة أو المتقطعة أو المتميزة discrete المعدودة counted

التكرار والتشتت والتنوع

- أولا: التكرار
  - مفاهيم أساسية: التكرار المطلق، والتكرار النسبي، وأساس المعيار (في كل حجم معين)
  - لتمثيل التكرار عدة أساليب بعد معالجتها. من أفضل الأساليب: خط الانحدار أو خط التوافق الأفضل

	التكرار										
الصفة	699	695	656	599	570	567	565	553	552	542	508
الفعل	2062	1976	2043	1995	2241	2056	2233	2276	2056	2089	2339

- هل هناك علاقة بين استخدام الصفات والأفعال؟ الجواب: نعم، والعلاقة تناسبية عكسية (كلما زاد استخدام المؤلفين للصفات قل استخدامهم للأفعال).



التكرار والتشتت والتنوع

- **التكرار: التوزيع الطبيعي وغير الطبيعي**
- **القيمة الشاذة**
- **قياس النزعة المركزية**
- **القياسات الإحصائية: المتوسط، والوسيط، والمدى، والانحراف المعياري ...[عمليات أولية بسيطة] والنمذجات كالانحدار الخطي والانحدار اللوجستي (نمذجة التأثيرات الممتزجة mixed-effects models)**
- **الاختبار الإحصائي: إحصائيات تتجاوز العينة للاستدلال على شيء يتعلق بالمجموع اللغوي في العينة: (مثلا: كاي تربيع وحجم التأثير)**
- **حجم التأثير (مثال لحالة):**

قيمة الارتباط الأسماء وعدد الصفات (ر =52.، مجال الثقة 95% ([.46, .58]) وهي ليست كبيرة كقيمة الارتباط بين الأفعال والضمائر (ر =81.، مجال الثقة 95% ([.775, .836]) التي تفسر تواردهما في ثلثي الحالات

التكرار والتشتت والتنوع

- التشتت. مثال لتبسيط فكرة التشتت، جمع بيانات لغوية وتهيئتها لأجل معرفة مدى تشتت استعمال (السعادة) في التغريدات حسب عدد المستخدمين بين المناطق الإدارية في المملكة العربية السعودية.
- التشتت يخرجنا من مشكلة Whelk: (الاعتماد على التكرار وحدة يُعد مضللا دائما، وتشير هذه المشكلة إلى التوزيع غير المتكافئ). ما الحل المقترح؟ لنفترض الآتي:

كلمة السعادة في التواصل الاجتماعي	الجزء1	الجزء2	الجزء3	الجزء4	الجزء5	الجزء6
مجموع العينة الخام (الكلمات)	100,000	100,000	200,000	200,000	200,000	200,000
تكرار الكلمة	10	4	2	0	24	10
التكرار النسبي للكلمة في كل 100 ألف كلمة	10	2	1	0	12	5
هل تشتمل على الكلمة؟	نعم	نعم	نعم	لا	نعم	نعم

- باستخدام توزيع Juilland's D: نستخلص المتوسط (5) ثم الانحراف المعياري (4.55)، ويُقسم الانحراف المعياري على المتوسط: (الناتج: 0.91). يلي ذلك استخراج الحد الأقصى لقيمة التنوع حسب عدد أجزاء المنطقة (الجذر التربيعي لعدد أجزاء المنطقة − 1)

$$\text{توزيع جولاند} = 1 - \frac{0.91}{\sqrt{6-1}} = 0.59$$

- الصفر= التوزيع متباين للغاية وَ (1) = التنويع متساوٍ للغاية

التكرار والتشتت والتنوع

استخدام التكرارات النسبية لكلمة لتفسير حقيقة أحجام العينات غير المتساوية

- التكرار النسبية (لكل مليون كلمة) لكلمة السعادة

الفترة	2006	2007	2008	2009	2010
التكرار النسبي	1473.69	1609.75	1623.78	1505.46	1283.21

- الفروق بين التكرارات النسبية

الفترة	2006/2007	2007/2008	2008/2009	2009/2010
القيمة1– القيمة2	136.06	14.03	118.32–	222.25–

- التكرار النسبي ذو اللوغاريتم المحول (لكل مليون)

الفترة	2006	2007	2008	2009	2010
log2	10.53	10.65	10.67	10.56	10.33

علم الدلالة والمعاني

- قياس التوارد العشوائي (التكرارات الملحوظة والمتوقعة): تقييم لقوة معاني الكلمات من عدم القوة
- القياسات في علم الدلالة والمعاني للكلمات تتطلب معالجة للتصاحب اللفظي
  - التكرار النسبي
  - المعلومات المتبادلة
  - الأرجحية اللوغارثمية
  - قيمة–ز
  - قيمة–ت
  - الدايس
  - اللوج دايس
  - نسبة اللوج
  - الفعالية ذات النهاية الصغرى
  - دلتا ب
  - كوهن–د

النحو المعجمي (الانحدار اللوجستي)

	نوع_السياق	نوع_الاسم	طول_المركب.الاسمي	النوع (التعريف والتنكير)
1	أ_غير.محدد	د_عَلَم	29	ب_تعريف
2	ب_محدد	ب_معدود_جمع	15	ب_تعريف
3	ب_محدد	ج_غير.معدود	7	ب_تعريف
4	ب_محدد	ب_معدود_جمع	7	ب_تعريف
5	ب_محدد	ج_غير.معدود	27	ب_تعريف
6	ب_محدد	أ_معدود_مفرد	14	ب_تعريف
7	أ_غير.محدد	أ_معدود_مفرد	4	أ_تنكير
8	ب_محدد	د_عَلَم	17	ب_تعريف
9	أ_غير.محدد	أ_معدود_مفرد	51	أ_تنكير
10	ب_محدد	أ_معدود_مفرد	26	ب_تعريف
11	ب_محدد	أ_معدود_مفرد	3	ب_تعريف
12	ب_محدد	أ_معدود_مفرد	4	ب_تعريف
13	ب_محدد	أ_معدود_مفرد	6	ب_تعريف

>التمييز بين المنطلق وبين القيم الأخرى للمتغيرات الفئوية (الاسمية والترتيبية): أ_ (المنطلق)، وبقية الترميز للمتغيرات الفئوية

>المتنبئات: نوع السياق، ونوع الاسم، وطول المركب الاسمي

>**المتغير الناتج: التعريف والتنكير**

>المتغير الناتج (المتغير اللغوي مثل: الخصائص النحوية) ومتغير المتنبئ (المتغير المستقل مثل: المتغيرات السياقية المرتبطة بتلك الخصائص)

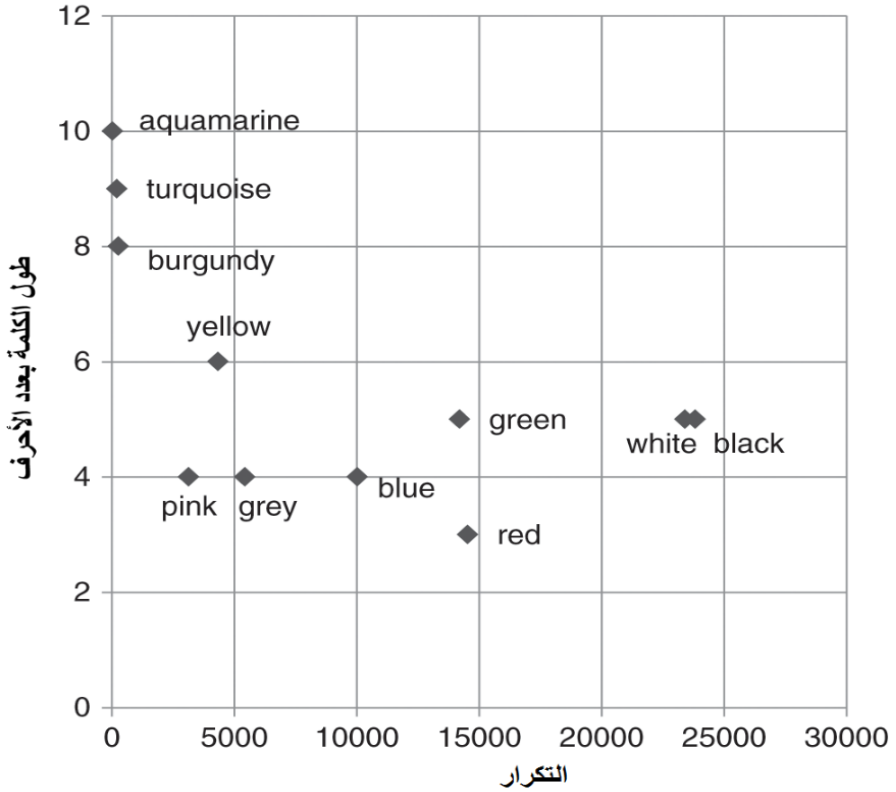
بناء النموذج ثم اختيار الأكثر دقة

اسم النموذج	الناتج	المتنبئات المتضمَّنة	النتيجة بناءً على ناتج البرنامج الإحصائي
المنطلق	النوع	[اسم]	نموذج المنطلق مع الحصر فقط
✓ الأول	النوع	نوع_السياق	ذو دلالة إحصائية، أي: أفضل بكثير من نموذج المنطلق معيار معلومات أكايكي = 30.86
الثاني	النوع	نوع_السياق نوع_الاسم	أخطاء المعيار الكبيرة -> حدث خطأ ما يُسبّب نوع_الاسم مشكلة مع الفصل الكامل.
الثالث	النوع	نوع_السياق طول_المركب.الاسمي	ذو دلالة إحصائية، أي: أفضل بكثير من نموذج المنطلق؛ ومع ذلك، ليس أفضل بكثير من النموذج الأول معيار معلومات أكايكي = 31.91 (أكبر من معيار معلومات أكايكي للنموذج الأول)

	التقدير (الأرجحية اللوغارثمية)	خطأ المعيار
(الحصر intercept)	-21.056	4530.376
نوع_السياق.ب_محدد	23.889	4530.376
نوع_السياق.ب_معدود_جمع	18.733	6706.381
نوع_الاسم.ج_غير.معدود	18.733	9244.108
نوع_الاسم.ج_غير.معدود	39.961	8958.692

	التقدير للأرجحية اللوغارثمية	خطأ المعيار	قيمة ز (اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	95% مجال الثقة أقل	95% مجال الثقة أعلى
(الحصر intercept)	-3.219	1.020	-3.156	0.002	0.04	0.002	0.189
نوع_السياق.ب_محدد	6.802	1.247	5.457	0.000	900	116.878	21421.229
طول_المركب.الاسمي	0.037	0.039	0.939	0.348	1.037	0.966	1.138

التنوع اللغوي: التجميع clustering



1. إحداثية التكرار وطول الجملة لـ green وblue [14205, 5] و[10035, 4] على التوالي.

2. يتطلب منا اختصار المسافات بتحويل القيم إلى $z\text{-scores}_2$

3. قيمة-z_2 = $\frac{\text{القيمة-المتوسط}}{\text{الانحراف المعياري(العينة)}}$ (تختلف عن قيمة z z-score)

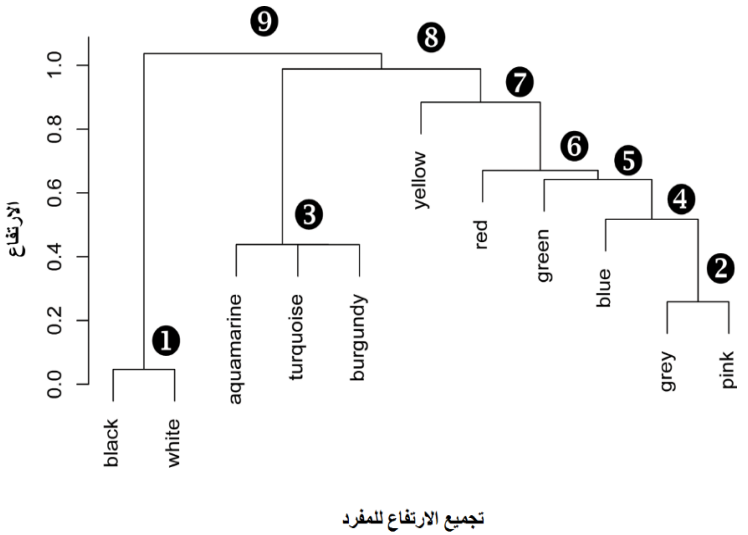
4. هناك: تحتاج إلى استخراج المتوسط، ثم استخراج الانحراف المعياري [الجذر التربيعي لمجموع مربعات انحراف القيم عن المتوسط مقسومة على عدد القيم-1

$$\text{الانحراف المعياري}_{\text{للعينة}} = \sqrt{\frac{\text{مجموع مربعات المسافات من المتوسط}}{\text{مجموع عدد أجزاء المدونة-1}}}$$

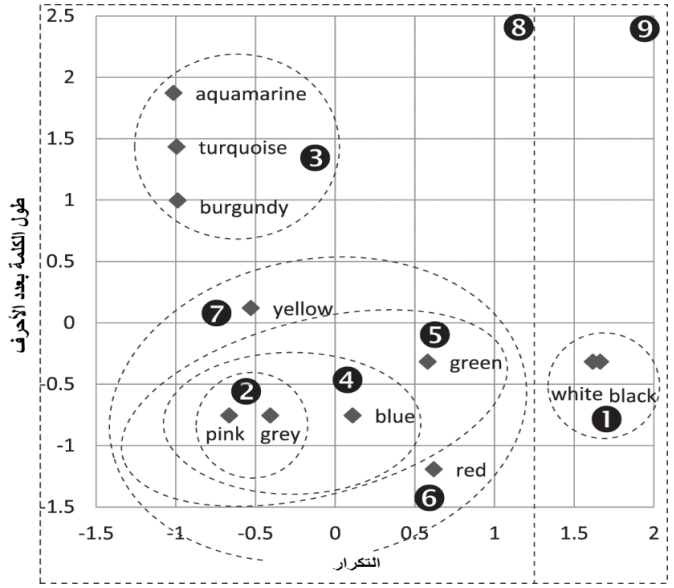
التنوع اللغوي: التجميع clustering

إحداثية التكرار وطول الجملة لـ green و blue [14205, 5] و[10035, 4] على التوالي

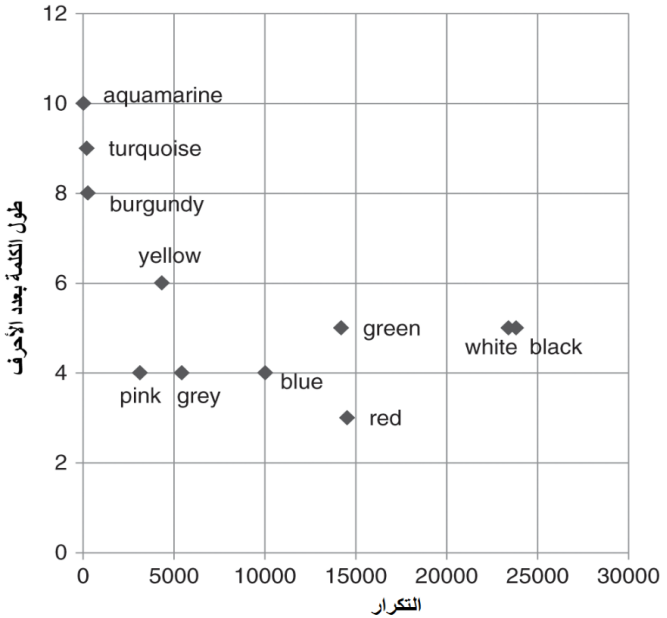
إحداثيات قيمة–ز2 الجديدة لـ green و blue [0.58, –0.32] و[0.11, –0.76] على التوالي



مخطط تفرّعي لطريقة الربط المنفرد



التجميع



إحداثية التكرار وطول الكلمة

المسافة الإقليدية: $\sqrt{(0.11 - 0.58)^2} + [-0.32 - (-0.76)]^2 = 0.64$

مسافة مانهاتن: $|0.11 - 0.58| + |-0.32 - (-0.76)| = 0.91$

البيانات اللغوية التاريخية: تحليل التقلب الاستعمال

تقلب الاستعمال للكلمة

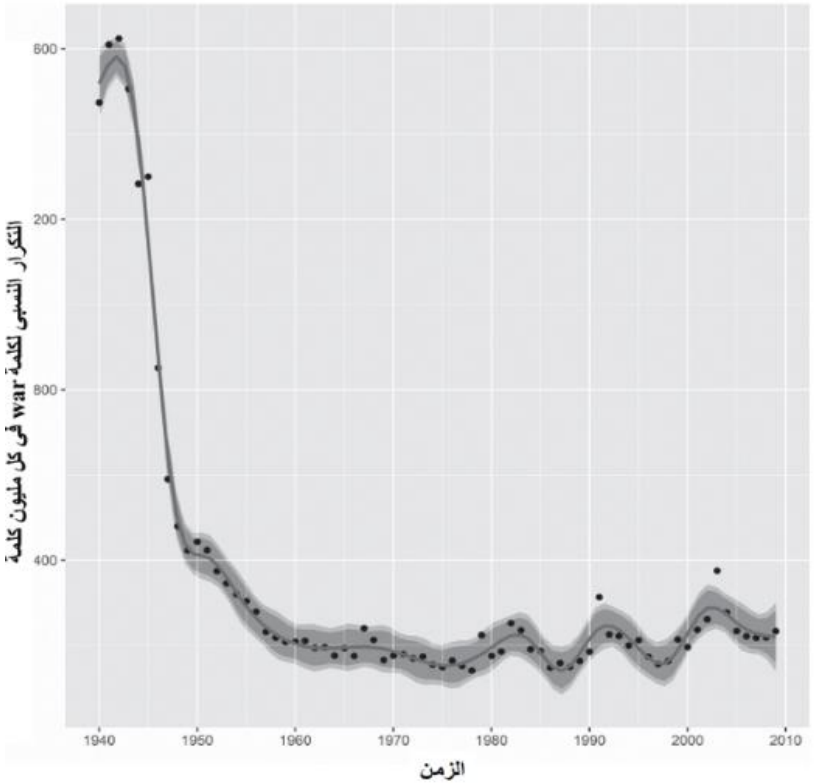
تمثيل نتائج تحليل التقلب الاستعمالي UFA على البيانات اللغوية التاريخية (كلمة الحرب مثلا)

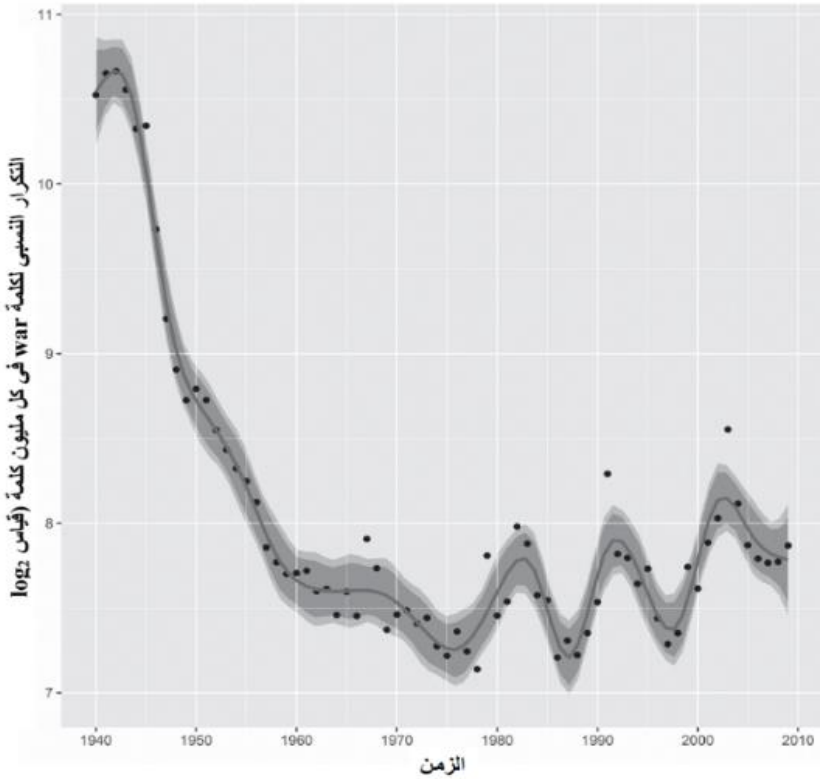
القمم والقعور للكلمة والمتصاحبات

الهدف: تصنيف المتصاحبات على أنها متسقة consistent، أو مستجدة initiating، أو مهجورة terminating، أو متحولة transient

الإجراء: تمثيل نقاط الاختلاف (نقاط ذات قيم معاملات المواءمة الدني [نسبة عدد الموافقة على عدد الحالات كنقاط اختلاف للمعنى للمتصاحبات) بآلية القمم والقعور peaks and troughs

البيانات اللغوية التاريخية: تحليل التقلب الاستعمال





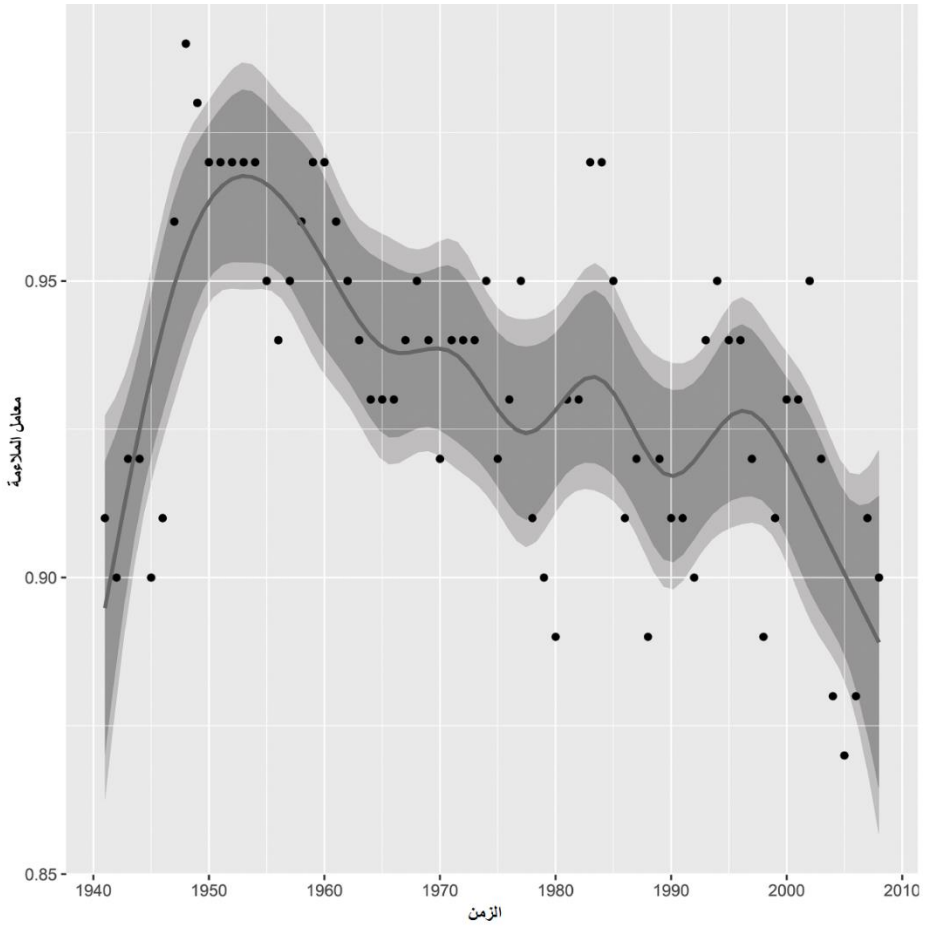
تحليل التقلب الاستعمالي UFA على البيانات اللغوية التاريخية

- تمثيل لحالة (الحرب)

x = السنوات

y = التكرار النسبي

البيانات اللغوية التاريخية: تحليل التقلب الاستعمال



القمم والقعور للكلمة والمتصاحبات

- تمثيل لحالة (الحرب ومتصاحباتها)

x = السنوات

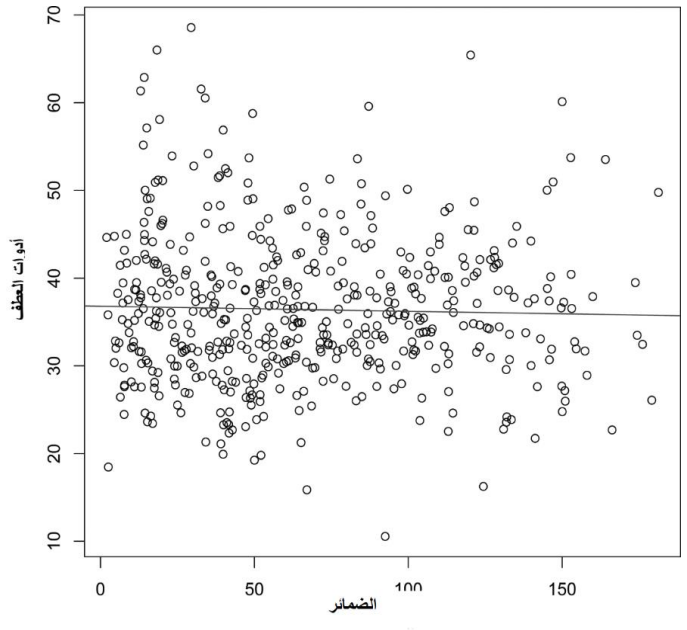
y = معامل الملاءمة أو المواءمة

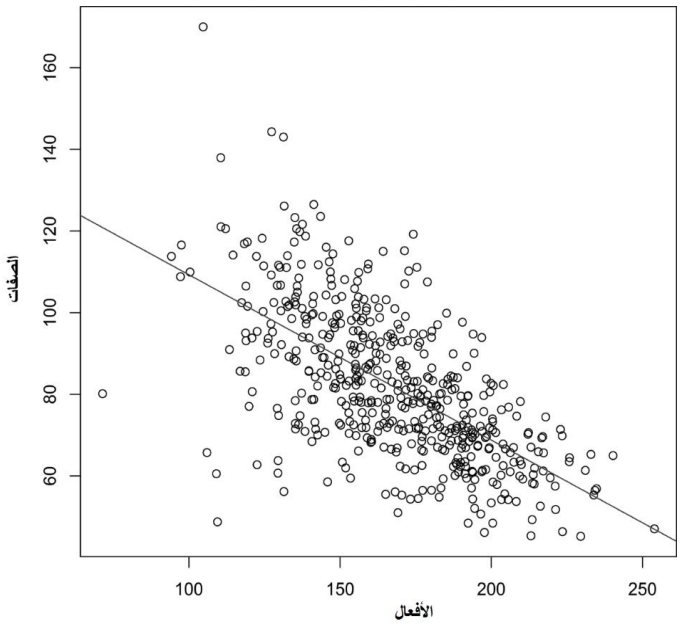
التنوع اللغوي: الارتباط

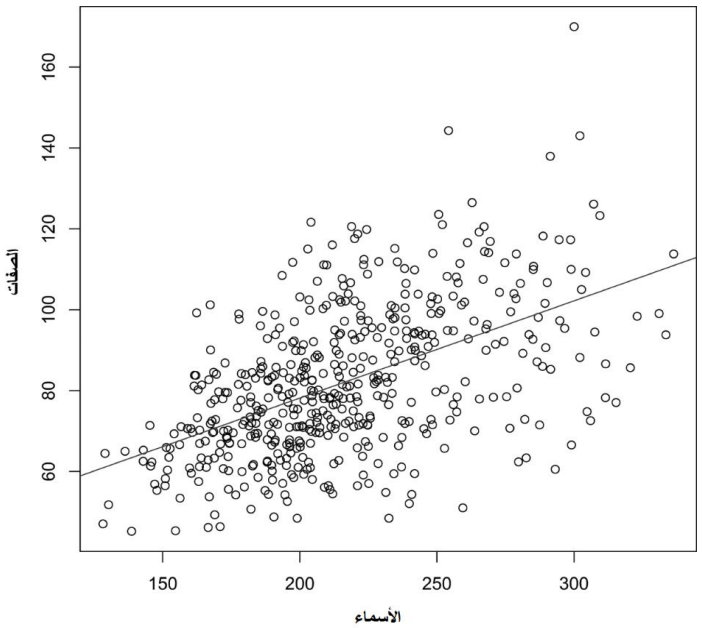
- هل بين متغيرين لغويين ارتباط؟
- إذا كان لدينا عشر ملاحظات فقط من نصوص مكتوبة أو حوارات فنحن بحاجة إلى ارتباط كبير قدره (0.63) للوصول إلى الدلالة الإحصائية
- في حالة وجود 100 ملاحظة يكون المزيد من التغير مقبولًا وتنخفض القيمة الحرجة الضرورية لارتباط بيرسون إلى (0.2)
- في حالة وجود 1000 ملاحظة سيكون هناك ارتباط ضئيل بقيمة (0.06) وذو دلالة إحصائية
- من المنظور اللغوي فإن للارتباط بقيمة 0.06 (أو0.2 هنا) تأثيرًا عمليًا ضئيلًا للغاية من جهة قوة العلاقة بين المتغيرات
- أيا كانت القيمة؛ لابد من الإبلاغ عنها

التنوع اللغوي: الارتباط

- هل بين متغيرين لغويين ارتباط؟

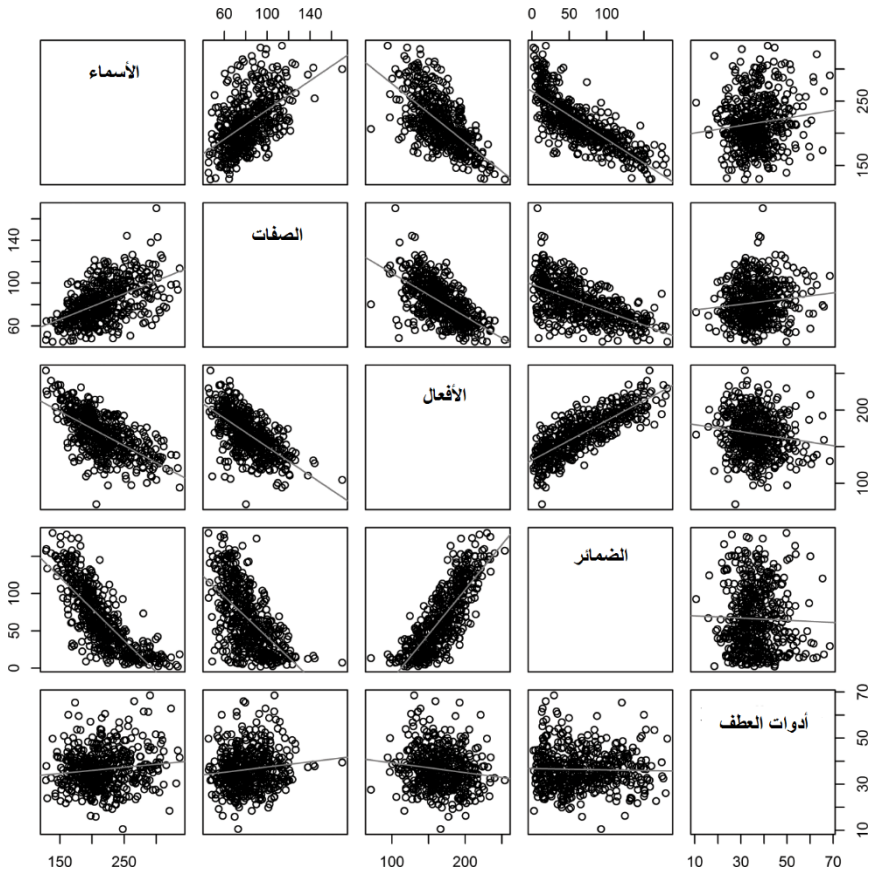




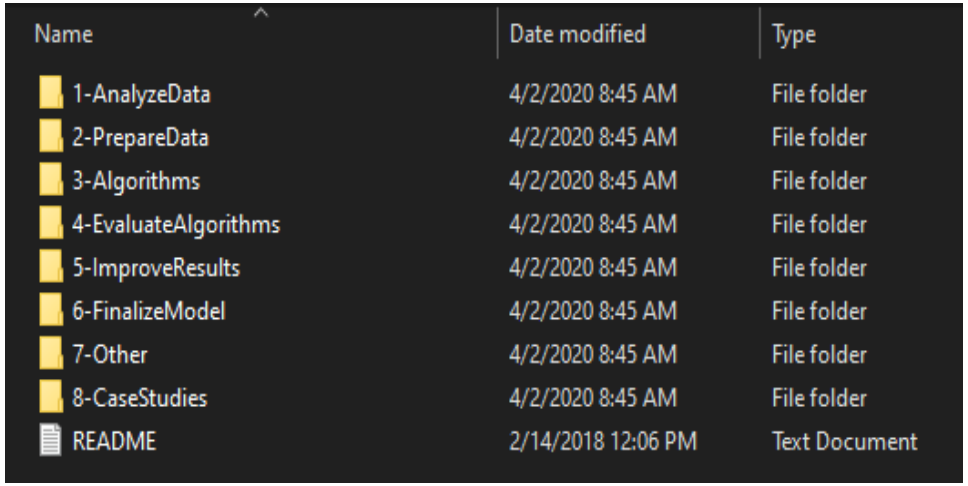


التنوع اللغوي: الارتباط

- الارتباط: مخطط التبعثر بين الأسماء والصفات والأفعال والضمائر والأدوات



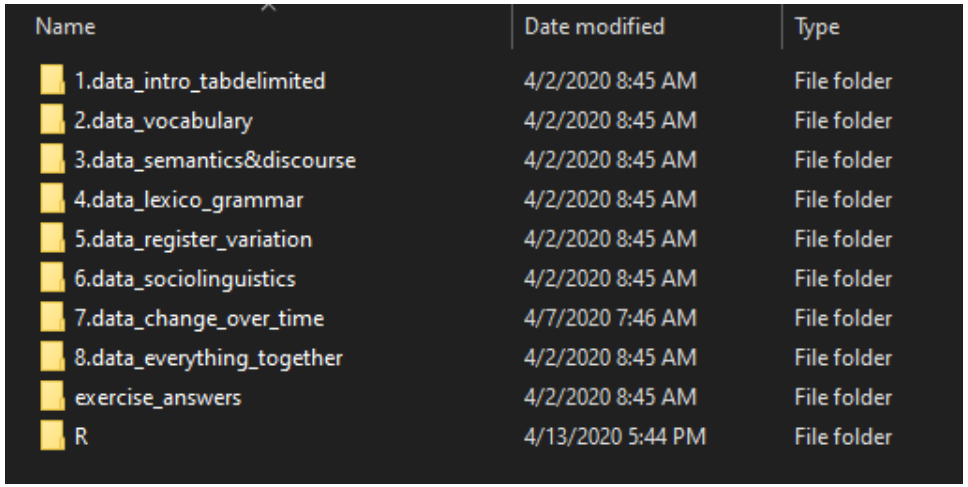
ملفات التدريب



Name	Date modified	Type
1-AnalyzeData	4/2/2020 8:45 AM	File folder
2-PrepareData	4/2/2020 8:45 AM	File folder
3-Algorithms	4/2/2020 8:45 AM	File folder
4-EvaluateAlgorithms	4/2/2020 8:45 AM	File folder
5-ImproveResults	4/2/2020 8:45 AM	File folder
6-FinalizeModel	4/2/2020 8:45 AM	File folder
7-Other	4/2/2020 8:45 AM	File folder
8-CaseStudies	4/2/2020 8:45 AM	File folder
README	2/14/2018 12:06 PM	Text Document

- التعرف على آر والمكتبات (تحليل البيانات، وتهيئة البيانات، والخوارزميات وتقويمها، وتحسين النتائج، واختيار النموذج، وتحسين أداء النموذج، ودراسة الحالات [البدء في مشروع من البداية إلى النهاية] المجلد [9].

https://machinelearningmastery.com/



Name	Date modified	Type
1.data_intro_tabdelimited	4/2/2020 8:45 AM	File folder
2.data_vocabulary	4/2/2020 8:45 AM	File folder
3.data_semantics&discourse	4/2/2020 8:45 AM	File folder
4.data_lexico_grammar	4/2/2020 8:45 AM	File folder
5.data_register_variation	4/2/2020 8:45 AM	File folder
6.data_sociolinguistics	4/2/2020 8:45 AM	File folder
7.data_change_over_time	4/7/2020 7:46 AM	File folder
8.data_everything_together	4/2/2020 8:45 AM	File folder
exercise_answers	4/2/2020 8:45 AM	File folder
R	4/13/2020 5:44 PM	File folder

- الإحصاءات الوصفية والاستدلالية وتحليل التكرار والشيوع والخطاب والنحو المعجمي والتنوع واللغويات الاجتماعية والنفسية والتاريخية [8-1]

نهاية العرض

شكرًا لاستماعكم

للتواصل: salmujaiwel@hotmail.com