تحليل المشاعر

متوسط | تحليل نصوص | Orange Data Mining | KNIME Analytics Platform

وصف المشكلة

تصنيف مراجعات عربية قصيرة إلى إيجابية أو سلبية، ومقارنة تمثيل الكلمات بتمثيل مقاطع الحروف (character n-grams) المناسب للغة العربية الغنية صرفياً.

الأهداف

- تحويل النص إلى خصائص رقمية
- فهم أثر التمثيل على الأداء
- قراءة النتائج بواقعية مع بيانات صغيرة

البيانات

المجموعة	الحجم	المصدر	الترخيص
مراجعات عربية قصيرة	160 صفاً × 2 عموداً، نحو 9.1 كيلوبايت	الأكاديمية (نصوص أصلية مؤلفة للتعليم)	ترخيص CC BY 4.0 من أكاديمية الذكاء الاصطناعي بدون برمجة (dlarabic.com)

الخوارزميات

بايز الساذج (Naive Bayes)، الانحدار اللوجستي (Logistic Regression)

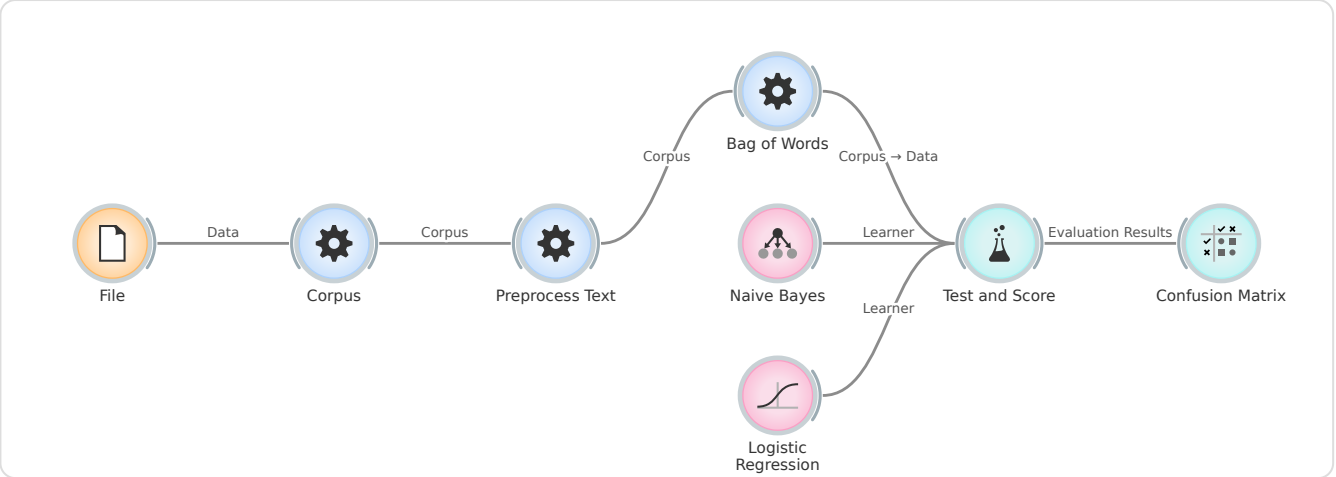
خطوات التنفيذ

1. في KNIME: File ← Import KNIME Workflow... واختر sentiment-analysis.knwf (بعد تثبيت إضافة KNIME Textprocessing)، ثم Execute all.
2. افتح sentiment-analysis.ows في Orange بعد تثبيت إضافة Text (Options ← Add-ons).
3. في Orange (بعد تثبيت إضافة Text): Test and Score ← Bag of Words ← Preprocess Text ← Corpus مع Naive Bayes وLogistic Regression.
4. في KNIME (إضافة Text Processing): Bag of Words Creator ← Document Vector ← Category to Class (يعيد عمود الصنف بعد Document Vector) ← Preprocessing ← Strings to Document ← CSV Reader ← Learner.

5. جرّب تمثيل المقاطع الحرفية إن توفر، وقارن.

مخطط سير العمل

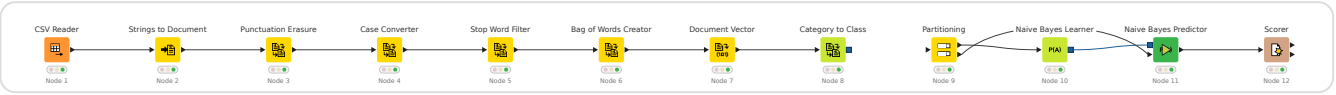
Orange



مرسوم من ملف سير العمل بصيغة OWS المرفق في الحزمة، بأيقونات العناصر الأصلية وألوان فئاتها وأسماء القنوات.

1. File → Corpus (Data)
2. Corpus → Preprocess Text (Corpus)
3. Preprocess Text → Bag of Words (Corpus)
4. Bag of Words → Test and Score (Corpus → Data)
5. Naive Bayes → Test and Score (Learner)
6. Logistic Regression → Test and Score (Learner)
7. Test and Score → Confusion Matrix (Evaluation Results)

KNIME



مخطط بأسماء عُقد KNIME وأيقوناتها الأصلية وألوان أنواعها، مطابق لملف KNWF المرفق في الحزمة (نُفّذ بنجاح في محرك KNIME 5.12).

1. CSV Reader → Strings to Document
2. Strings to Document → Punctuation Erasure
3. Punctuation Erasure → Case Converter
4. Case Converter → Stop Word Filter
5. Stop Word Filter → Bag of Words Creator
6. Bag of Words Creator → Document Vector
7. Document Vector → Category to Class
8. Category to Class → Partitioning
9. Partitioning → Naive Bayes Learner
10. Naive Bayes Learner → Naive Bayes Predictor
11. Naive Bayes Predictor → Scorer
12. Partitioning → Naive Bayes Predictor

الملفات

الملف	المسار داخل الحزمة / الرابط
مجموعة البيانات CSV	arabic-reviews.csv

ملف WEKA بصيغة ARFF	arabic-reviews.arff
سير عمل Orange (.ows)	sentiment-analysis.ows
البيانات بصيغة Orange (.tab) لسير العمل	arabic-reviews.tab
النتائج المرجعية المحسوبة (CSV)	reviews-results.csv
سير عمل KNIME (.knwf): نُفّذ بنجاح في محرك KNIME 5.12 (لم يُفتح في الواجهة)	sentiment-analysis.knwf

ملف KNIME (.knwf) بُني على KNIME 5.12 ونُفّذت كل عُقده بنجاح في محرك KNIME دون فتح الواجهة؛ افتحه من File ← Import KNIME Workflow.

النتائج

نتائج فعلية محسوبة على الملفات المرفقة

مرجع scikit-learn 1.9.1، 5-fold stratified CV, shuffle, random_state=42:

- Naive Bayes (word BoW): Accuracy 50.6% (± 5.0%), F1-macro 0.502
- Logistic Regression (word BoW): Accuracy 54.4% (± 4.2%), F1-macro 0.541
- Naive Bayes (char 2-4 grams): Accuracy 56.2% (± 4.0%), F1-macro 0.560
- Logistic Regression (char 2-4 grams): Accuracy 69.4% (± 7.5%), F1-macro 0.690

لم نوفر سير عمل Orange لهذا المشروع لأنه يحتاج إضافة Text غير المثبتة افتراضياً.

مقاييس التقييم

- Accuracy
- F1-macro
- Confusion Matrix

تفسير النتائج

بـ160 مراجعة فقط يبقى تمثيل الكلمات قريباً من التخمين (نحو 50–54%) لأن معظم الكلمات تظهر مرة واحدة، بينما رفعت المقاطع الحرفية الأداء إلى نحو 69% لأنها تلتقط الجذور والسوابق واللواحق. الدرس: حجم البيانات والتمثيل أهم من اختيار الخوارزمية.

أسئلة للمناقشة

- لماذا تفيد المقاطع الحرفية في العربية تحديداً؟
- كيف تتعامل مع النفي مثل «غير جيد»؟

- كم مراجعة تقدّر أننا نحتاج لنتيجة موثوقة؟

البيانات الاصطناعية في هذه الحزمة مولّدة بالبرمجة لأغراض تعليمية (موسومة بكلمة synthetic) وليست بيانات حقيقية. الأرقام في قسم النتائج محسوبة فعلياً على الملفات المرفقة فقط حيث يُذكر ذلك. — dlarabic.com