HW2b – Big Data and ML Systems

YL5090

Task: Text classification based on TF-IDF and Word2Vec

20NewsGroups contains emails from 20 news groups. I first split the data set into a train set and a test set with 9:1 ratios. After training the data on the training set, I evaluate the overall prediction accuracy, average precision and recall for each model + preprocessing method. I also evaluate the precision and recall individually for each group for each model + preprocessing method.

Preprocessing:

1. Transform data into Spark SQL DataFrame
2. Removing punctuations, numbers, and transfer to lowercases
3. Remove stopwords and tokenize

Models tested:

* Logistic Regression with TFIDF and Word2Vec feature selections
* Naïve Bayes with TFIDF, as Naïve Bayes requires nonnegative inputs so Word2Vec is discarded
* RandomForest with TFIDF and Word2Vec feature selections

Models not used:

* Gradient Boosted Tree: Currently ML only supports binary classification. Thus discarded.
* SVM: computationally intensive so not used this time.
* DecisionTree: because random forest is an ensembling method it should be better than decision tree.

Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Avg Precision | Avg Recall |
| Naïve Bayes + TFIDF | 0.696 | 0.698 | 0.690 |
| Logistic Regression + TFIDF | 0.047 | NA | 0.05 |
| Logistic Regression + W2V | 0.592 | 0.624 | 0.576 |
| Random Forest + TFIDF | 0.586 | 0.584 | 0.573 |
| Random Forest + W2V | 0.586 | 0.584 | 0.573 |

The best performing algorithm for TFIDF weighting is Naïve Bayes, while the best performing algorithm for Word2Vec representation is Logistic Regression, in this case.

The results listed above: Logistic Regression + TFIDF produces suspiciously low accuracy, and might need further checking. Random Forest produces the same results for different document representation, thus might need further checking too.

Taking only the results for Naïve Bayes + TFIDF and Logistic Regression + Word2Vec.

Insights from the data:

* Some of the groups that have the best prediction precision and recalls when using word2vec are:
  + Alt.atheism
  + Rec.sports.baseball
  + Rec.sports.hockey
  + Rec.sports.crypt
  + Soc.religion.christian
  + Talk.politics.misc
  + Talk.politics.mideast
* Some of the groups that have the best prediction precision and recalls when using tfidf are:
  + Alt.atheism
  + Rec.sports.baseball
  + Rec.sports.hockey
  + Rec.sports.crypt
  + Soc.religion.christian
  + Talk.politics.guns
  + Talk.politics.misc

These groups are more or less “easy to classify”. Since we represent document in its collection of words, these groups likely have their distinctive word occurrence patterns. For example, it’s likely the word “God” occurs pretty often in alt.atheism group and other religion related groups, but not in other groups. For sports, it's likely the emails have many game-related terms. As for politics, when it is more specific e.g. about guns or a specific region, the classification accuracy is higher.

A general insight is thus, the more “narrow” the topic is, the more likely the group will have a distinct set of vocabulary, the more likely classification accuracy will be higher for that group.

Error analysis:

|  |  |  |
| --- | --- | --- |
| Topic | LR W2V Precision | LR W2V Recall |
| alt.atheism (17) | 0.886 | 0.736 |
| comp.graphics  (13.0) | 0.593 | 0.525 |
| comp.os.ms-windows.misc(9.0) | 0.623 | 0.559 |
| comp.sys.ibm.pc.hardware(11.0) | 0.468 | 0.712 |
| comp.sys.mac.hardware(14.0) | 0.583 | 0.264 |
| comp.windows.x (8.0) | 0.551 | 0.796 |
| misc.forsale (12.0) | 0.607 | 0.627 |
| rec.autos (6.0) | 0.637 | 0.394 |
| rec.motorcycles (2.0) | 0.482 | 0.768 |
| rec.sport.baseball (3.0) | 0.716 | 0.696 |
| rec.sport.hockey (0.0) | 0.807 | 0.893 |
| sci.crypt (4.0) | 0.742 | 0.821 |
| sci.electronics (10.0) | 0.553 | 0.426 |
| sci.med (5.0) | 0.737 | 0.412 |
| sci.space( 7.0) | 0.477 | 0.564 |
| soc.religion.christian (1.0) | 0.479 | 0.817 |
| talk.politics.guns (16.0) | 0.429 | 0.526 |
| talk.politics.mideast (15.0) | 0.568 | 0.778 |
| talk.politics.misc (18.0) | 0.909 | 0.200 |
| talk.religion.misc (19.0) | NA | 0.000 |

|  |  |  |
| --- | --- | --- |
| Topic | LR TFIDF Precision | LR TFIDF Recall |
| alt.atheism (17) | NA | 0 |
| comp.graphics  (13.0) | NA | 0 |
| comp.os.ms-windows.misc(9.0) | NA | 0 |
| comp.sys.ibm.pc.hardware(11.0) | NA | 0 |
| comp.sys.mac.hardware(14.0) | NA | 0 |
| comp.windows.x (8.0) | NA | 0 |
| misc.forsale (12.0) | NA | 0 |
| rec.autos (6.0) | NA | 0 |
| rec.motorcycles (2.0) | NA | 0 |
| rec.sport.baseball (3.0) | NA | 0 |
| rec.sport.hockey (0.0) | NA | 0 |
| sci.crypt (4.0) | 0.047 | 1 |
| sci.electronics (10.0) | NA | 0 |
| sci.med (5.0) | NA | 0 |
| sci.space( 7.0) | NA | 0 |
| soc.religion.christian (1.0) | NA | 0 |
| talk.politics.guns (16.0) | NA | 0 |
| talk.politics.mideast (15.0) | NA | 0 |
| talk.politics.misc (18.0) | NA | 0 |
| talk.religion.misc (19.0) | NA | 0 |

|  |  |  |
| --- | --- | --- |
| Topic | NB TFIDF Precision | NB TFIDF Recall |
| alt.atheism (17) | 0.672 | 0.811 |
| comp.graphics  (13.0) | 0.5 | 0.705 |
| comp.os.ms-windows.misc(9.0) | 0.667 | 0.034 |
| comp.sys.ibm.pc.hardware(11.0) | 0.615 | 0.769 |
| comp.sys.mac.hardware(14.0) | 0.538 | 0.660 |
| comp.windows.x (8.0) | 0.761 | 0.648 |
| misc.forsale (12.0) | 0.661 | 0.661 |
| rec.autos (6.0) | 0.679 | 0.746 |
| rec.motorcycles (2.0) | 0.654 | 0.768 |
| rec.sport.baseball (3.0) | 0.846 | 0.797 |
| rec.sport.hockey (0.0) | 0.913 | 0.84 |
| sci.crypt (4.0) | 0.849 | 0.804 |
| sci.electronics (10.0) | 0.603 | 0.623 |
| sci.med (5.0) | 0.769 | 0.735 |
| sci.space( 7.0) | 0.645 | 0.727 |
| soc.religion.christian (1.0) | 0.810 | 0.662 |
| talk.politics.guns (16.0) | 0.726 | 0.789 |
| talk.politics.mideast (15.0) | 0.882 | 0.833 |
| talk.politics.misc (18.0) | 0.643 | 0.72 |
| talk.religion.misc (19.0) | 0.529 | 0.474 |

|  |  |  |
| --- | --- | --- |
| Topic | RF W2V Precision | RF W2V Recall |
| alt.atheism (17) | 0.780 | 0.736 |
| comp.graphics  (13.0) | 0.540 | 0.442 |
| comp.os.ms-windows.misc(9.0) | 0.438 | 0.476 |
| comp.sys.ibm.pc.hardware(11.0) | 0.474 | 0.519 |
| comp.sys.mac.hardware(14.0) | 0.533 | 0.453 |
| comp.windows.x (8.0) | 0.557 | 0.630 |
| misc.forsale (12.0) | 0.618 | 0.356 |
| rec.autos (6.0) | 0.452 | 0.535 |
| rec.motorcycles (2.0) | 0.577 | 0.653 |
| rec.sport.baseball (3.0) | 0.774 | 0.696 |
| rec.sport.hockey (0.0) | 0.831 | 0.853 |
| sci.crypt (4.0) | 0.707 | 0.732 |
| sci.electronics (10.0) | 0.468 | 0.475 |
| sci.med (5.0) | 0.512 | 0.618 |
| sci.space( 7.0) | 0.487 | 0.691 |
| soc.religion.christian (1.0) | 0.657 | 0.648 |
| talk.politics.guns (16.0) | 0.565 | 0.614 |
| talk.politics.mideast (15.0) | 0.644 | 0.704 |
| talk.politics.misc (18.0) | 0.632 | 0.48 |
| talk.religion.misc (19.0) | 0.429 | 0.158 |

|  |  |  |
| --- | --- | --- |
| Topic | RF TFIDF Precision | RF TFIDF Recall |
| alt.atheism (17) | 0.780 | 0.736 |
| comp.graphics  (13.0) | 0.540 | 0.442 |
| comp.os.ms-windows.misc(9.0) | 0.438 | 0.476 |
| comp.sys.ibm.pc.hardware(11.0) | 0.474 | 0.519 |
| comp.sys.mac.hardware(14.0) | 0.533 | 0.453 |
| comp.windows.x (8.0) | 0.557 | 0.630 |
| misc.forsale (12.0) | 0.618 | 0.356 |
| rec.autos (6.0) | 0.452 | 0.535 |
| rec.motorcycles (2.0) | 0.577 | 0.653 |
| rec.sport.baseball (3.0) | 0.774 | 0.696 |
| rec.sport.hockey (0.0) | 0.831 | 0.853 |
| sci.crypt (4.0) | 0.707 | 0.732 |
| sci.electronics (10.0) | 0.468 | 0.475 |
| sci.med (5.0) | 0.512 | 0.618 |
| sci.space( 7.0) | 0.487 | 0.691 |
| soc.religion.christian (1.0) | 0.657 | 0.648 |
| talk.politics.guns (16.0) | 0.565 | 0.614 |
| talk.politics.mideast (15.0) | 0.644 | 0.704 |
| talk.politics.misc (18.0) | 0.632 | 0.48 |
| talk.religion.misc (19.0) | 0.429 | 0.158 |