# Logistic Regression Model for Survey Open Response Document Classification

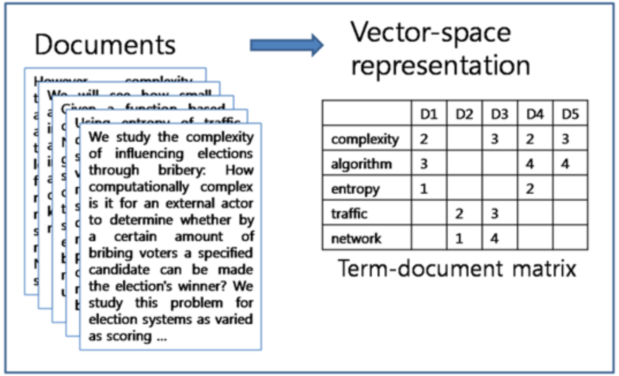
by Lee Werner

**Key words**: natural language processing, logistic regression, python, survey open responses, tf-idf, machine learning, regression model, document vectorization

BE THE MATCH®

# Flow chart

*Vectorize the survey open responses (documents) using TF-IDF*



We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

| | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|
| complexity | 2 | | 3 | 2 | 3 |
| algorithm | 3 | | | 4 | 4 |
| entropy | 1 | | | 2 | |
| traffic | | 2 | 3 | | |
| network | | 1 | 4 | | |

Term-document matrix

https://insights.dice.com/2015/03/16/how-we-data-mine-related-tech-skills/

*Train regression model using vectorized documents*

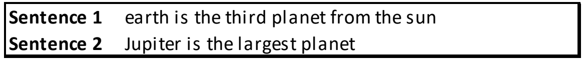
$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

*Make predictions on the test set and measure accuracy*

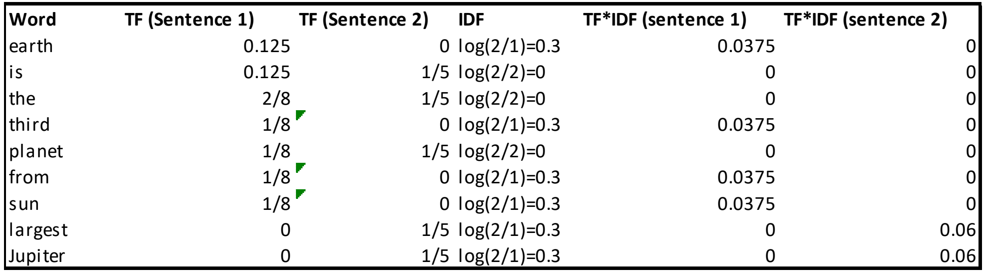
| Responses | Predicted | Actual |
|---|---|---|
| i am over 60 years old so it looks like i cant help with this right joekessyahoocom | 2 | 2 |
| i'm not going to pay 100 to join | 3 | 3 |
| i am above the age restriction cut off | 2 | 2 |
| i don't think i'll match anyone | 14 | 14 |
| my addressid information is going to change soon and i dont want to fill this out before im more stable | 14 | 14 |
| im in canada had to go through canadian blood services apologies for my mistake | 7 | 7 |
| i'm in canada and it won't allow me to put in my province or postal code | 7 | 7 |
| my cousin died from this and i am scared | 1 | 8 |
| i didn't comply with the medical selection as i have arthritis in both ankles | 1 | 1 |

# Document Vectorization using TF-IDF

- Converts documents to a matrix representation that takes term uniqueness into account.
- Analyzes how frequently a term appears on a document (term-frequency/TF) and compares it with how often it is expected to appear on an average page (inverse document frequency/IDF).



| | |
|---|---|
| **Sentence 1** | earth is the third planet from the sun |
| **Sentence 2** | Jupiter is the largest planet |



| Word | TF (Sentence 1) | TF (Sentence 2) | IDF | TF*IDF (sentence 1) | TF*IDF (sentence 2) |
|---|---|---|---|---|---|
| earth | 0.125 | 0 | log(2/1)=0.3 | 0.0375 | 0 |
| is | 0.125 | 1/5 | log(2/2)=0 | 0 | 0 |
| the | 2/8 | 1/5 | log(2/2)=0 | 0 | 0 |
| third | 1/8 | 0 | log(2/1)=0.3 | 0.0375 | 0 |
| planet | 1/8 | 1/5 | log(2/2)=0 | 0 | 0 |
| from | 1/8 | 0 | log(2/1)=0.3 | 0.0375 | 0 |
| sun | 1/8 | 0 | log(2/1)=0.3 | 0.0375 | 0 |
| largest | 0 | 1/5 | log(2/1)=0.3 | 0 | 0.06 |
| Jupiter | 0 | 1/5 | log(2/1)=0.3 | 0 | 0.06 |

https://blog.expertrec.com/what-is-tf-idf-and-how-does-it-work/

$$w_{i,j} = tf_{i,j} * \log \frac{N}{df_i}$$

- $w^{i,j}$ = tf-idf value
- $tf^{i,j}$= number of occurences of *i* in *j*
- $df^{i}$ =number of documents containing *i*
- *N* = total number of documents

*i* = word (and/or word pair)
*j* = document

https://inspiremelabs.com/tf-idf-for-seo/

# Logistic Regression Model

**Equation:**

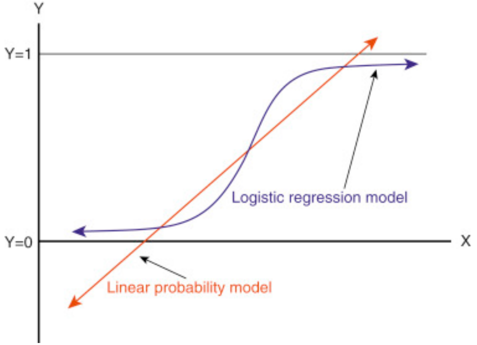
$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

- *a* is the intercept
- *b* is the slope
- *P* is the probability (0 to 1) that the target variable *X* belongs to a particular category, where *X* is a document in the form of a TF-IDF vector

## Logistic Regression Model Properties

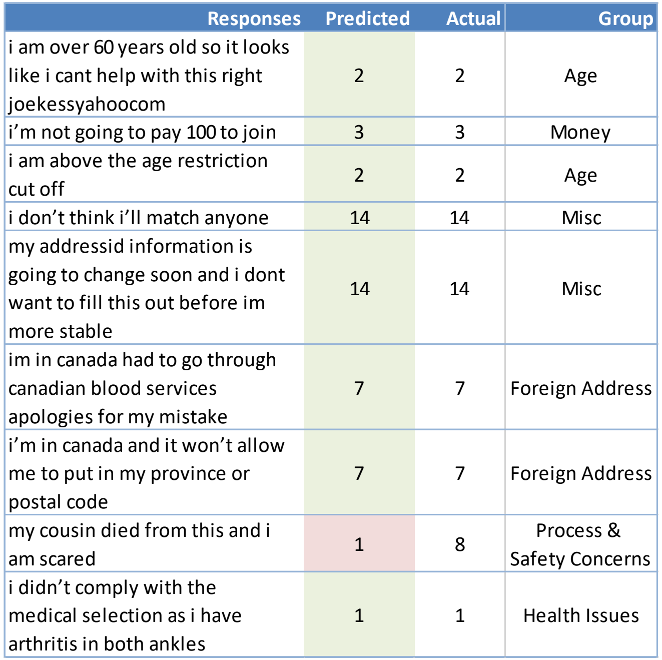
- Target is discrete (binary or ordinal)
- Predicted values are the probability of a particular level(s) of the target variable at the given values of the input variables

https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5



https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis

# Results



| Responses | Predicted | Actual | Group |
|---|---|---|---|
| i am over 60 years old so it looks like i cant help with this right joekessyahoocom | 2 | 2 | Age |
| i’m not going to pay 100 to join | 3 | 3 | Money |
| i am above the age restriction cut off | 2 | 2 | Age |
| i don’t think i’ll match anyone | 14 | 14 | Misc |
| my addressid information is going to change soon and i dont want to fill this out before im more stable | 14 | 14 | Misc |
| im in canada had to go through canadian blood services apologies for my mistake | 7 | 7 | Foreign Address |
| i’m in canada and it won’t allow me to put in my province or postal code | 7 | 7 | Foreign Address |
| my cousin died from this and i am scared | 1 | 8 | Process & Safety Concerns |
| i didn’t comply with the medical selection as i have arthritis in both ankles | 1 | 1 | Health Issues |

- Accuracy
  - The model correctly predicted 88.45% of document responses in the test set (n=407).
- Looking ahead
  - This model can be used for *any* survey that is routinely used and that has open-response questions.
  - The model used for this presentation had a training set of only n=1628. With more training documents the model’s accuracy will increase.