# sa_pos_neg

May 31, 2016

## 1 Text data analysis with Knowledge-based system

## 2 Data preperation

We will use a dataset consisting of baby product reviews on Amazon.com.

```
In [1]: NUMBER_OF_REVIEWS_TO_ANALYZE = 100000
```

```
In [2]: import pandas as pd
```

```
In [3]: products = pd.read_csv("../valt_sa_data/amazon_baby.csv")[['review', 'ratin
```

```
In [4]: products = products[0:NUMBER_OF_REVIEWS_TO_ANALYZE]
```

```
In [5]: products
```

```
Out[5]:                                                   review  rating
        0      These flannel wipes are OK, but in my opinion ...       3
        1      it came early and was not disappointed. i love...       5
        2      Very soft and comfortable and warmer than it l...       5
        3      This is a product well worth the purchase.  I ...       5
        4      All of my kids have cried non-stop when I trie...       5
        5      When the Binky Fairy came to our house, we did...       5
        6      Lovely book, it's bound tightly so you may not...       4
        7      Perfect for new parents. We were able to keep ...       5
        8      A friend of mine pinned this product on Pinter...       5
        9      This has been an easy way for my nanny to reco...       4
        10     I love this journal and our nanny uses it ever...       4
        11     This book is perfect!  I'm a first time new mo...       5
        12     I originally just gave the nanny a pad of pape...       4
        13     I thought keeping a simple handwritten journal...       3
        14     Space for monthly photos, info and a lot of us...       5
        15     I bought this calender for myself for my secon...       4
        16     I love this little calender, you can keep trac...       5
        17     This was the only calender I could find for th...       5
        18     I completed a calendar for my son's first year...       4
        19     We wanted to get something to keep track of ou...       5
```

```
20     I had a hard time finding a second year calend...       5
21     I only purchased a second-year calendar for my...       2
22     I LOVE this calendar for recording events of m...       5
23     Calendar is exactly as described, but I find t...       3
24     Wife loves this calender. Comes with a lot of ...       5
25     My daughter had her 1st baby over a year ago. ...       5
26     Extremely useful! As a new mom, tired and inex...       5
27     My son loves peek a boo at this age of 9 month...       3
28     One of baby's first and favorite books, and it...       4
29     I like how the book has a hook to attach it to...       5
...                                                  ...     ...
99970  We had an earlier version of this cup for our ...       2
99971  It's hard to tell from the picture, but the sp...       4
99972  this is a good beaker except for that the spou...       4
99973  High quality step stool. My tot often stands r...       5
99974  I bought this to mainly help my toddler reach ...       5
99975                          Fits well. Soft to touch.       5
99976  I love these paci's which are so helpful to my...       1
99977  I recommend this training cup. I have bought M...       5
99978  I just love MAM products. High quality and pri...       5
99979  We love this bottle.  Mam bottles are great, a...       5
99980  I bought three of these, in addition to some o...       4
99981  Not as easy to remove as other brands, like Ro...       2
99982  After about a year and some time more, I'm tak...       5
99983  I love this wrap.  My husband I both use it da...       5
99984  I bought this for my daughter and she loves it...       5
99985  Love this wrap. Cotton, soft and light. Inexpe...       5
99986  I love this wrap I purchased the large in colo...       5
99987  This is a wonderful wrap to carry baby or todd...       5
99988  Overall, I am pleased with this purchase.  I w...       4
99989  I loved this bag! I didn't like the color but ...       4
99990  I spent a surprising amount of time searching ...       5
99991  I have 4 Bumble Bags. The quality is top-notch...       5
99992  ONLY WISH IT HAD A SPACE FOR THE PULL TIGHTENE...       4
99993  Boy car seat covers are expensive so I settled...       5
99994  The Snuzzler is perfect for my baby boy. It ma...       5
99995  This is excellent padding especially for your ...       5
99996  Despite what it says, you may not use this in ...       2
99997  Bought this to protect the leather seats in ou...       5
99998  After doing my research online, I found this t...       5
99999  We had a similar product for our first car sea...       5

[100000 rows x 2 columns]
```

## 2.1 Build the word count vector for each review

Let us explore a specific example of a baby product.

```
In [6]: products.iloc[9]
```

```
Out[6]: review    This has been an easy way for my nanny to reco...
        rating                                                   4
        Name: 9, dtype: object
```

Now, we will perform 2 simple data transformations:

1. Remove punctuation using Python's built-in string functionality.
2. Transform the reviews into word-counts of positive and negative word.
3. Finally made prediction based on positive/negative words ratio.

```
In [7]: def remove_punctuation(text):
            import string
            return text.translate(None, string.punctuation)

        review_without_puctuation = products['review'].apply(str).apply(remove_punc
```

```
In [8]: significant_words = pd.read_csv('../valt_sa_data/positive-negative-words.cs

        positive_words = pd.read_csv('../valt_sa_data/positive-words.csv', header=N
        negative_words = pd.read_csv('../valt_sa_data/negative-words.csv', header=N

        def count_number_of_significant_words(text):
            prediction = 3
            words = text['review'].split()
            word_dict = {}
            for word in significant_words:
                word_dict[word] = 0
            for word in words:
                if word in significant_words:
                    if word not in word_dict:
                        word_dict[word] = 1
                    else:
                        word_dict[word] = word_dict[word] + 1
            positive = 0
            negative = 0
            for positive_word in positive_words:
                if positive_word in word_dict:
                    positive += word_dict[positive_word]

            for negative_word in negative_words:
                if negative_word in word_dict:
                    negative += word_dict[negative_word]

            n = positive + negative
            if n > 0:
                prediction = 1 + int(round(float(positive) / n * 4))
```

```
    return pd.Series(prediction)

predictions_df = pd.DataFrame(review_without_puctuation).apply(count_number
predictions_df.columns = ['prediction']

products_with_words = products.join(predictions_df)
```

Now, let us see what rating and predictions look like.

```
In [9]: products_with_words

Out[9]:                                                   review  rating  predictio
        0      These flannel wipes are OK, but in my opinion ...       3
        1      it came early and was not disappointed. i love...       5
        2      Very soft and comfortable and warmer than it l...       5
        3      This is a product well worth the purchase.  I ...       5
        4      All of my kids have cried non-stop when I trie...       5
        5      When the Binky Fairy came to our house, we did...       5
        6      Lovely book, it's bound tightly so you may not...       4
        7      Perfect for new parents. We were able to keep ...       5
        8      A friend of mine pinned this product on Pinter...       5
        9      This has been an easy way for my nanny to reco...       4
        10     I love this journal and our nanny uses it ever...       4
        11     This book is perfect!  I'm a first time new mo...       5
        12     I originally just gave the nanny a pad of pape...       4
        13     I thought keeping a simple handwritten journal...       3
        14     Space for monthly photos, info and a lot of us...       5
        15     I bought this calender for myself for my secon...       4
        16     I love this little calender, you can keep trac...       5
        17     This was the only calender I could find for th...       5
        18     I completed a calendar for my son's first year...       4
        19     We wanted to get something to keep track of ou...       5
        20     I had a hard time finding a second year calend...       5
        21     I only purchased a second-year calendar for my...       2
        22     I LOVE this calendar for recording events of m...       5
        23     Calendar is exactly as described, but I find t...       3
        24     Wife loves this calender. Comes with a lot of ...       5
        25     My daughter had her 1st baby over a year ago. ...       5
        26     Extremely useful! As a new mom, tired and inex...       5
        27     My son loves peek a boo at this age of 9 month...       3
        28     One of baby's first and favorite books, and it...       4
        29     I like how the book has a hook to attach it to...       5
        ...                                                  ...     ...         ..
        99970  We had an earlier version of this cup for our ...       2
        99971  It's hard to tell from the picture, but the sp...       4
        99972  this is a good beaker except for that the spou...       4
        99973  High quality step stool. My tot often stands r...       5
        99974  I bought this to mainly help my toddler reach ...       5
```

```
99975                            Fits well. Soft to touch.       5
99976  I love these paci's which are so helpful to my...       1
99977  I recommend this training cup. I have bought M...       5
99978  I just love MAM products. High quality and pri...       5
99979  We love this bottle.  Mam bottles are great, a...       5
99980  I bought three of these, in addition to some o...       4
99981  Not as easy to remove as other brands, like Ro...       2
99982  After about a year and some time more, I'm tak...       5
99983  I love this wrap.  My husband I both use it da...       5
99984  I bought this for my daughter and she loves it...       5
99985  Love this wrap. Cotton, soft and light. Inexpe...       5
99986  I love this wrap I purchased the large in colo...       5
99987  This is a wonderful wrap to carry baby or todd...       5
99988  Overall, I am pleased with this purchase.  I w...       4
99989  I loved this bag! I didn't like the color but ...       4
99990  I spent a surprising amount of time searching ...       5
99991  I have 4 Bumble Bags. The quality is top-notch...       5
99992  ONLY WISH IT HAD A SPACE FOR THE PULL TIGHTENE...       4
99993  Boy car seat covers are expensive so I settled...       5
99994  The Snuzzler is perfect for my baby boy. It ma...       5
99995  This is excellent padding especially for your ...       5
99996  Despite what it says, you may not use this in ...       2
99997  Bought this to protect the leather seats in ou...       5
99998  After doing my research online, I found this t...       5
99999  We had a similar product for our first car sea...       5

[100000 rows x 3 columns]
```

## 2.2 Evaluate the model

```
In [10]: y_true = products_with_words['rating']
         y_predicted = products_with_words['prediction']

         from sklearn.metrics import confusion_matrix
         cm = confusion_matrix(y_true, y_predicted)

         print 'Confusion matrix:'
         print cm

         from sklearn.metrics import classification_report

         print 'Classification report:'
         print classification_report(y_true, y_predicted)
```

```
Confusion matrix:
[[ 1169  1818  3353  1683   897]
 [  413   807  2345  1762   997]
 [  381   707  2850  3015  2204]
 [  296   736  4224  7058  5977]
 [  707  1296  9185 18505 27615]]
Classification report:
             precision    recall  f1-score   support

          1       0.39      0.13      0.20      8920
          2       0.15      0.13      0.14      6324
          3       0.13      0.31      0.18      9157
          4       0.22      0.39      0.28     18291
          5       0.73      0.48      0.58     57308

avg / total       0.52      0.39      0.43    100000
```