Weather and Text for Investment Research

Kyle Balkissoon
June 2018
Kyle.Balkissoon@ibm.com

Predictive Models are easier to improve by adding meaningful features vs complex statistics (or more quants)

Bloomberg
FACTSET
THOMSON REUTERS

+

+

The Weather Company
An IBM Business

Standard Financial Data

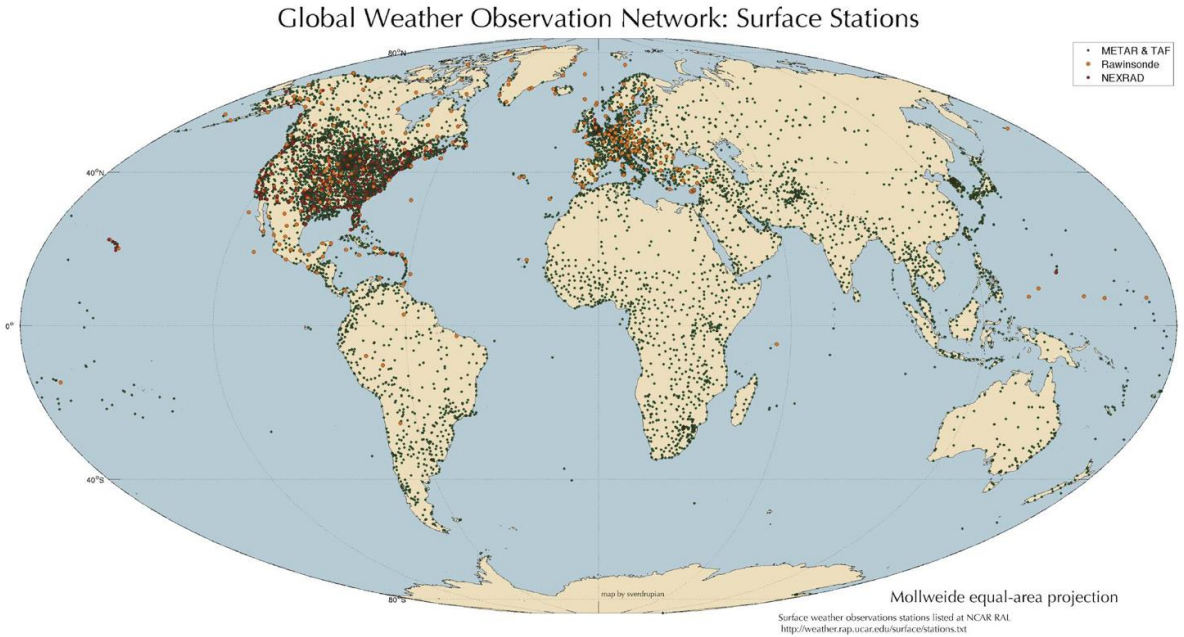
Quant Researchers

- Low Marginal Lift
- Demands free food
- Too many GARCH discussions

More Data

- Potentially high lift
- No food demands
- Does not talk about GARCH

Weather Stations Globally – Capturing Data

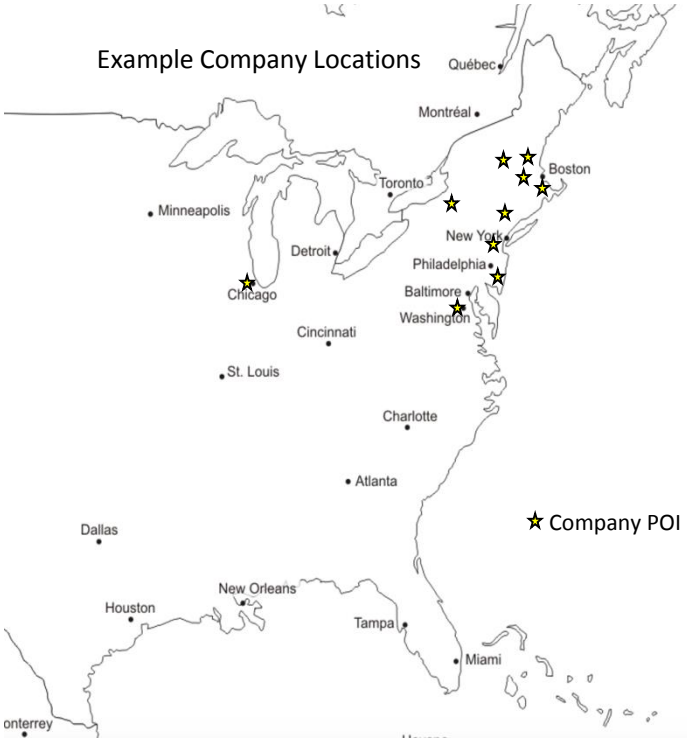


Weather Data

```
> head(data)
   key zipcode       class expire_time_gmt obs_id obs_name valid_time_gmt day_ind temp wx_icon icon_extd     wx_phrase pressure_tend  pressure_desc dewPt heat_index rh pressure vis wc wdir
1 KSTL      NA observation      1483260660   KSTL St Louis     1483253460       N   27      29      2900 Partly Cloudy            NA                 21         27 78    29.43  10 27   NA
2 KSTL      NA observation      1483264260   KSTL St Louis     1483257060       N   28      29      2900 Partly Cloudy            NA                 21         28 75    29.43  10 28   NA
3 KSTL      NA observation      1483267860   KSTL St Louis     1483260660       N   27      26      2600        Cloudy             1         Rising    20         27 75    29.43  10 27   NA
4 KSTL      NA observation      1483271460   KSTL St Louis     1483264260       N   28      26      2600        Cloudy            NA                 21         28 75    29.43   7 28   NA
5 KSTL      NA observation      1483275060   KSTL St Louis     1483267860       N   27      26      2600        Cloudy            NA                 20         27 75    29.43   8 27   NA
6 KSTL      NA observation      1483278660   KSTL St Louis     1483271460       N   25      26      2600        Cloudy             1 Rising Rapidly    20         25 81    29.44   7 25   70
  wdir_cardinal gust wspd max_temp min_temp precip_total precip_hrly snow_hrly uv_desc feels_like uv_index qualifier qualifier_svrty blunt_phrase terse_phrase secondary_swell_height
1          CALM   NA    0       NA       NA           NA           0       NA     Low         27        0       NA              NA           NA           NA                     NA
2          CALM   NA    0       NA       NA           NA           0       NA     Low         28        0       NA              NA           NA           NA                     NA
3          CALM   NA    0       NA       NA           NA           0       NA     Low         27        0       NA              NA           NA           NA                     NA
4          CALM   NA    0       NA       NA           NA           0       NA     Low         28        0       NA              NA           NA           NA                     NA
5          CALM   NA    0       NA       NA           NA           0       NA     Low         27        0       NA              NA           NA           NA                     NA
6           ENE   NA    3       NA       NA           NA           0       NA     Low         25        0       NA              NA           NA           NA                     NA
  primary_swell_direction clds primary_swell_height secondary_swell_direction primary_swell_period primary_wave_period water_temp secondary_swell_period primary_wave_height
1                      NA  CLR                   NA                        NA                   NA                 NA         NA                      NA                   NA
2                      NA  CLR                   NA                        NA                   NA                 NA         NA                      NA                   NA
3                      NA  CLR                   NA                        NA                   NA                 NA         NA                      NA                   NA
4                      NA  CLR                   NA                        NA                   NA                 NA         NA                      NA                   NA
5                      NA  CLR                   NA                        NA                   NA                 NA         NA                      NA                   NA
6                      NA  CLR                   NA                        NA                   NA                 NA         NA                      NA                   NA
>
```

Calculation of the Synthetic Weather Station – A Representation of weather temporally around a company of interest

- Calculation dependent on population surrounding company location
- If a company does not have POI data then weighting goes to METAR space



Equation

$$\text{Synthetic } Station = \frac{1}{\sum_{i=1}^{n} Population_i} * \frac{1}{n} \sum_{i=1}^{n} POI_{wx_i} * Population_i$$

$$where\ i\ is\ a\ POI\ location$$

Recommendation: use data.table for both mapping of POI to weather and aggregation

Aggregating Geospatial Weather Data to appropriate periods of interest

Synthetic Station Generator → **Feature Generator** → **Quarterly Aggregates**

Ticker Has POI data

Synthetic weather stations used to create weather factors around **each company** POI. Weighted by Population

Ticker Has NO POI data

Synthetic weather stations used to create weather factors around **all METAR Stations in key areas**. Weighted by Population

For Each Synthetic Station

- Departures from Normal
- Growing Degree Days
- Heating Degree Days
- Cooling Degree Days
- Business Hour Weather …

Weather Feature Normalization

Weather features normalized using anchored z-scores

$$z = \frac{x_i - \mu}{\sigma}$$

Average Over Period

$$WX_i = \frac{1}{n}\sum_{i=1}^{n} Var_i$$

Use data.table!!!

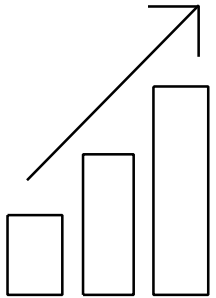
What formulating the prediction statement looks like:

Gathered weather data for commodity-growing regions

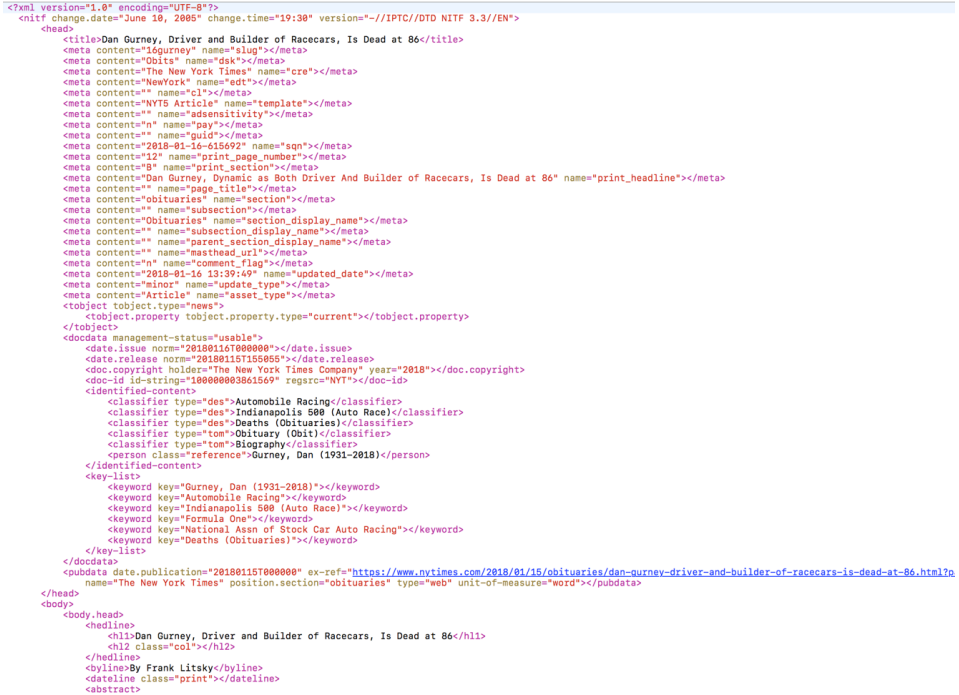
Built dynamic forecasting model using data and analyzed data using Machine Learning

$$TargetVariable_t = StandardPredictors_{t-1} + Precipitation_{t-1} + \ldots.$$

Identified new trading opportunities based on known features and weather resulting in a profitable strategy and a sharpe > 1



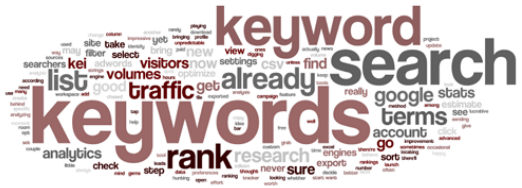
News Data: Raw News data in XML Format



```
<?xml version="1.0" encoding="UTF-8"?>
<nitf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
    <head>
        <title>Dan Gurney, Driver and Builder of Racecars, Is Dead at 86</title>
        <meta content="16gurney" name="slug"></meta>
        <meta content="Obits" name="dsk"></meta>
        <meta content="The New York Times" name="cre"></meta>
        <meta content="NewYork" name="edt"></meta>
        <meta content="" name="cl"></meta>
        <meta content="NYT5 Article" name="template"></meta>
        <meta content="" name="adsensitivity"></meta>
        <meta content="n" name="pay"></meta>
        <meta content="" name="guid"></meta>
        <meta content="2018-01-16-415692" name="sqn"></meta>
        <meta content="12" name="print_page_number"></meta>
        <meta content="B" name="print_section"></meta>
        <meta content="Dan Gurney, Dynamic as Both Driver And Builder of Racecars, Is Dead at 86" name="print_headline"></meta>
        <meta content="" name="page_title"></meta>
        <meta content="obituaries" name="section"></meta>
        <meta content="" name="subsection"></meta>
        <meta content="Obituaries" name="section_display_name"></meta>
        <meta content="" name="subsection_display_name"></meta>
        <meta content="" name="parent_section_display_name"></meta>
        <meta content="" name="masthead_url"></meta>
        <meta content="n" name="comment_flag"></meta>
        <meta content="2018-01-16 13:39:49" name="updated_date"></meta>
        <meta content="minor" name="update_type"></meta>
        <meta content="Article" name="asset_type"></meta>
        <tobject tobject.type="news">
            <tobject.property tobject.property.type="current"></tobject.property>
        </tobject>
        <docdata management-status="usable">
            <date.issue norm="20180116T000000"></date.issue>
            <date.release norm="20180115T155055"></date.release>
            <doc.copyright holder="The New York Times Company" year="2018"></doc.copyright>
            <doc-id id-string="100000003861569" regsrc="NYT"></doc-id>
            <identified-content>
                <classifier type="des">Automobile Racing</classifier>
                <classifier type="des">Indianapolis 500 (Auto Race)</classifier>
                <classifier type="des">Deaths (Obituaries)</classifier>
                <classifier type="tom">Obituary (Obit)</classifier>
                <classifier type="tom">Biography</classifier>
                <person class="reference">Gurney, Dan (1931-2018)</person>
            </identified-content>
            <key-list>
                <keyword key="Gurney, Dan (1931-2018)"></keyword>
                <keyword key="Automobile Racing"></keyword>
                <keyword key="Indianapolis 500 (Auto Race)"></keyword>
                <keyword key="Formula One"></keyword>
                <keyword key="National Assn of Stock Car Auto Racing"></keyword>
                <keyword key="Deaths (Obituaries)"></keyword>
            </key-list>
        </docdata>
        <pubdata date.publication="20180115T000000" ex-ref="https://www.nytimes.com/2018/01/15/obituaries/dan-gurney-driver-and-builder-of-racecars-is-dead-at-86.html?p
            name="The New York Times" position.section="obituaries" type="web" unit-of-measure="word"></pubdata>
    </head>
    <body>
        <body.head>
            <hedline>
                <hl1>Dan Gurney, Driver and Builder of Racecars, Is Dead at 86</hl1>
                <hl2 class="col"></hl2>
            </hedline>
            <byline>By Frank Litsky</byline>
            <dateline class="print"></dateline>
            <abstract>
```

End-to-end workflow for news data aggregation and analysis

Aggregation



Data is coming from **New York Times**

Have access to everything published (print and online) historically

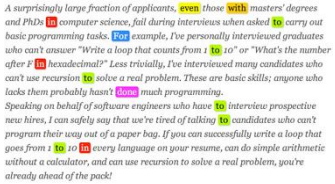
Requested articles from 2001 to the end of 2018. All articles include **full text** and **metadata** (section, date published, keywords, etc.)

Cleaning

All tickers are mapped to a **company name, nicknames,** and **subsidiary companies** (Darden to Olive Garden)

Found matches between company names and article text/keywords. Potential for articles to match to multiple companies

Analysis



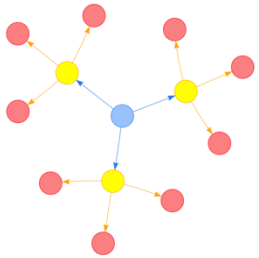
A surprisingly large fraction of applicants, even those with masters' degrees and PhDs in computer science, fail during interviews when asked to carry out basic programming tasks. For example, I've personally interviewed graduates who can't answer "Write a loop that counts from 1 to 10" or "What's the number after F in hexadecimal?" Less trivially, I've interviewed many candidates who can't use recursion to solve a real problem. These are basic skills; anyone who lacks them probably hasn't done much programming.
Speaking on behalf of software engineers who have to interview prospective new hires, I can safely say that we're tired of talking to candidates who can't program their way out of a paper bag. If you can successfully write a loop that goes from 1 to 10 in every language on your resume, can do simple arithmetic without a calculator, and can use recursion to solve a real problem, you're already ahead of the pack!

10+ features are extracted from article including: sentiment, taxonomy, and word count

R Packages: XML, tm
Sentiment: Dictionary Score/Proprietary

Feature creation and engineering for news data

Features

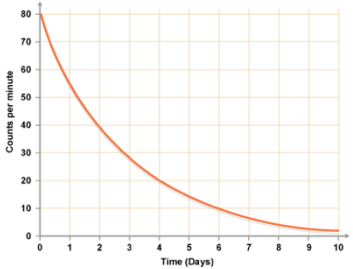


Example features that can be generated on a per taxonomy basis

1. Sentiment
2. Word Count
3. Rolling averages (looking backwards) using N day windows
4. Importance Score (combination of sentiment and word count)
5. Sentiment Index (3 Buckets)
6. Word Count Index (6 Buckets)

There are other features on a **daily basis** such as Sentiment and Word Count Index

Moving Forward



Decay algorithm is used to carry previous observations forward when no new news articles are released. Values decay at **X% rate** each day until a new article is released, resetting the value.

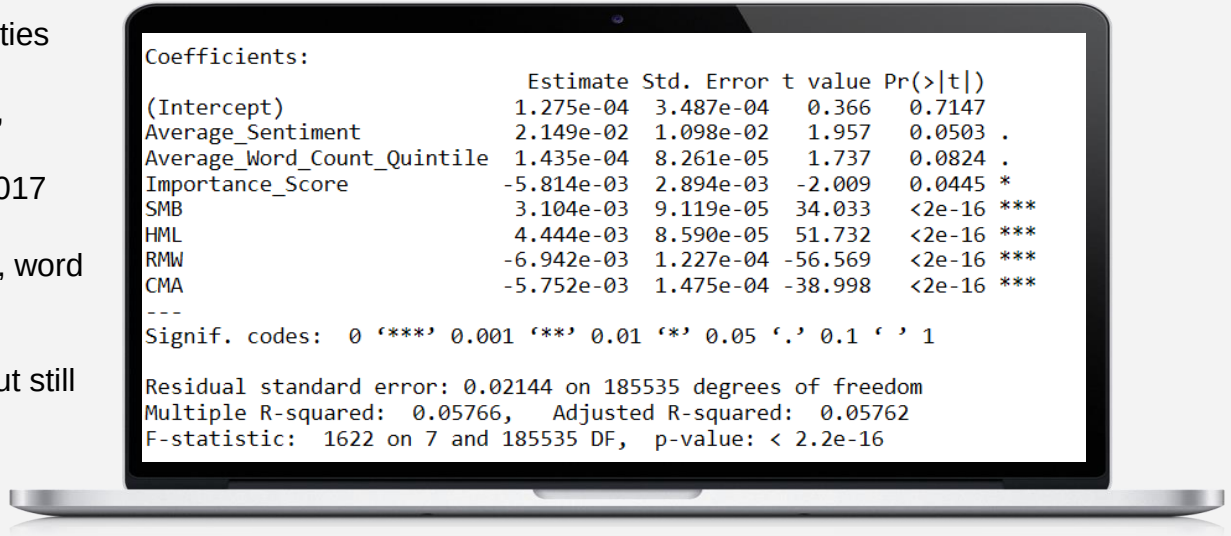
All values decay towards **zero**. For sentiment values that is trending towards **neutral** (1 is positive, -1 is negative). Word count trends towards zero from a positive number to show the **lessening importance/relevance** of an article over time.

Case Study: News & Equity Predictability

Problem Statement: Does news data have additive predictive power on equity returns?

$$Equity\ Returns = News\ Factors + Fama\ French\ Carhart\ Factors$$

- **Target Variable:** Daily Returns of 50 Equities Pool
- **Predictors**: News data (New York Times), Fama-French Carhart factors
- **Period:** January 2 2001– December 29 2017
- **Takeaways:**
  - News factors of average sentiment, word count and importance score are statistically significant
  - Not as significant as FFC factors but still provide some predictive power



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.275e-04  3.487e-04   0.366   0.7147
Average_Sentiment          2.149e-02  1.098e-02   1.957   0.0503 .
Average_Word_Count_Quintile 1.435e-04  8.261e-05   1.737   0.0824 .
Importance_Score           -5.814e-03  2.894e-03  -2.009   0.0445 *
SMB                         3.104e-03  9.119e-05  34.033   <2e-16 ***
HML                         4.444e-03  8.590e-05  51.732   <2e-16 ***
RMW                        -6.942e-03  1.227e-04 -56.569   <2e-16 ***
CMA                        -5.752e-03  1.475e-04 -38.998   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02144 on 185535 degrees of freedom
Multiple R-squared:  0.05766,   Adjusted R-squared:  0.05762
F-statistic:  1622 on 7 and 185535 DF,  p-value: < 2.2e-16
```



IBM