Quantifying Correlation between Financial News and Stocks

Haizhou Qu
Department of Computer Science, University of York
York, United Kingdom, YO105GH
hq524@york.ac.uk

Dimitar Kazakov
Department of Computer Science, University of York
York, United Kingdom, YO105GH
dimitar.kazakov@york.ac.uk

***Abstract*—Financial news and stocks appear linked to the point where the use of online news to forecast the markets has become a major selling point for some traders. The correlation between news content and stock returns is clearly of interest, but has been mostly centred on news meta-data, such as volume and popularity. We address this question here by measuring the correlation between the returns of 27 publicly traded companies and news about them as collected from Yahoo Financial News for the period 1 Oct 2014 to 30 Apr 2015. In all reported experiments, two metrics are defined, one to measure the distance between two time series, the other to quantify the difference between two collections of news items. Two 27 × 27 distance matrices are thus produced, and their correlation measured with the Mantel test. This allows us to estimate the correlation of stock market data (returns, change, volume and close price) with the content of published news in a given period of time. A number of representations for the news are tested, as well as different distance metrics between time series. Clear, statistically significant, moderate level correlations are detected in most cases. Lastly, the impact of the length of the period studied on the observed correlation is also investigated.**

I. Introduction

Using online financial news to assist forecasts using time series is very tempting as one's intuition suggests that there should be useful information in the news that is not directly reflected in the day-to-day figures reported for each publicly traded company. Indeed, there have been stock market traders, such as the now defunct Derwent Capital Markets hedge fund [1], which advertise that their forecasting algorithms make use of online social media, such as Twitter [2]. Nevertheless, any attempt to make use of such information brings up some difficult questions: What is the most useful representation of the text documents to be used? What features should one extract from them, and in what form? Can we decide when news is useful, if at all? Indeed, in addition to the implications of economic theories, such as the efficient market hypothesis, which seem to imply that the time series data available to traders only contains noise, one can also consider the case in which news only follows the markets with a certain delay, which would render it useless. These are hard questions and the road to answering them would be made easier if split into several stages.

We have previously looked at one particular financial event, the price crash of Volkswagen stock that followed the announcement of the US Environment Protection Agency investigation into what became known as the Dieselgate scandal. This was done in the hope that such an extreme event (where a substantial and sustained decline in price follows a news release) could provide an excellent data set on which to study the likely impact of financial news about a company on its performance on the stock market [3], with a focus on the potential causal link.

Here we change the perspective and instead want to study whether for a given time period, news and stock market data are correlated, leaving out the chicken-and-egg question of which one came first. In addition, we also ignore the exact time of news release, and combine all news about a given company published within the time period of interest into a single document. Then we study the differences between the news about a pair of companies, and how well such a difference is (co-)related to a difference in the performance of the two stocks over the same period. The chosen statistical measure, namely, the Mantel test, measures the correlation between differences in the news and in the time series for a whole set of companies at once, which should make the results less dependant on the circumstances of each individual company.

II. Data

We collected online news from Yahoo Financial News over the period 1 Oct 2014 – 30 Apr 2015. Each news item carries a time stamp (in EST time) and the symbols of one or more companies, to which it is related. The 27 stocks were selected to have no more than a total of 5 days with no news about them in the studied period. Very short news with less than 10 words or 100 characters were ignored, leaving a total of 67,840 news items.

We have also collected daily stock market data for the same companies and period of time. For each day and stock, the data set contains the $open$, $high$, low and $close$ price, as well as the $volume$ traded and the $adjusted\, close$ price. Table III shows a summary of the data available.

III. Method

A. The Mantel Test

The Mantel test is a statistical test to determine correlation between two pairwise distance matrices with the same rank [4].

Given two $n \times n$ distance matrices $\mathbf{U}$ and $\mathbf{V}$, it calculates the correlation r using equation 1.

$$r_{\mathbf{U},\mathbf{V}} = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{U_{ij} - \bar{u}}{\sigma_{\mathbf{U}}} \cdot \frac{V_{ij} - \bar{v}}{\sigma_{\mathbf{V}}} \tag{1}$$

where $m = n(n-1)/2$ is the number of pairwise distances of a size n population, $\bar{u}$ and $\bar{v}$ are means of pairwise distance elements located in upper triangle exclude the diagonal of $\mathbf{U}$ and $\mathbf{V}$.

For example, the matrix $\mathbf{U}$ may represent the genetic distances in a group of n species while another matrix $\mathbf{V}$ represents the geographic distance between the species' habitats. By applying the Mantel test, we can calculate how much the geographic distance between two species is correlated with their genetic differences. The idea behind the Mantel test is to randomly permute one matrix repeatedly while calculating each time a correlation according to equation 1. A hypothesis test is carried out to examine if this correlation is significantly lower than the one produced with the original matrices. Finding that this is the case would suggest a correlation between the two matrices, and the two sets of distances they represent. As already mentioned, we use the Mantel test here to determine the extent to which financial news and stocks are correlated with each other. In other words, we wanted to see whether the differences between a pair of time series, e.g. representing the daily returns of two stocks, are correlated with the differences in the news about these companies. For this purpose, it was necessary to define ways to measure the difference (or *distance*) between time series, and between text documents.

B. Comparing Time Series

We have considered three distance metrics to compare pairs of time series: the Cosine Distance (CD), the Euclidean Distance (ED) and the one produced by Pearson's correlation (further referred to as PD).

The well known cosine distance is calculated according to equation 2 where u_t and v_t are the values of each time series on day t.

Euclidean distance is determined by the length of the line segment connecting points $\mathbf{u}$ and $\mathbf{v}$ (see equation 3). It is a proper distance metric with wide-ranging spectrum of applications.

Pearson's correlation measures the linear correlation between a pair of variables (time series in our case) through the ratio of their covariance divided by the product of their standard deviations. Equation 4 shows how a distance metric can be defined on the basis of this correlation (see equation 4).

$$CD(\mathbf{u},\mathbf{v}) = 1 - \frac{\sum_t v_t \cdot u_t}{\sqrt{\sum_t u_t^2 \cdot \sum_t v_t^2}} \tag{2}$$

$$ED(\mathbf{u},\mathbf{v}) = \sqrt{\sum_t (u_t - v_t)^2} \tag{3}$$

$$PD(\mathbf{u},\mathbf{v}) = 1 - \frac{\sum_t (u_t - \bar{u})(v_t - \bar{v})}{\sqrt{\sum_t (u_t - \bar{u})^2 \sum_t (v_t - \bar{v})^2}} \tag{4}$$

where $\mathbf{u}$ and $\mathbf{v}$ are two vectors representing two time series with the same time index.

C. Comparing Texts

There is a number of representations developed for the purposes of Information Retrieval that could be used in this study. These range from the simplest bag-of-words model, which only takes into account the presence (and frequency) of words in a document, but ignores any word order, to representations of words and their neighbours (bigrams, trigrams, etc.) and those in which parts of the parse tree of a sentence are used as features [5].

A Bag-of-Words represents a collection of texts as a `document × word` matrix which treats each word in the whole collection as a separate feature. The content of each document is then encoded as a vector containing the (relative) frequency of each of its words, including zeros for all the words that do not appear in the document. This allows for an easy comparison between any two documents, at the price of ignoring the grammatical relationship between words. So, a set of text documents $\mathcal{D}$ is represented as a matrix M where each row corresponds a document $d \in \mathcal{D}$, and each column stands for a feature w (usually a word or token). Each element $M_{i,j}$ then is the relative frequency with which word j appears in document i.

An additional weighting scheme is often used to reduce the importance of words that appear across most documents, and highlight the ones that are characteristic to a small subset of documents. TF-IDF (Term Frequency – Inverse Document Frequency) [6] is the most popular such technique. Here the relative frequency of word w in document d is weighted according to equation 5. This reduces the perceived importance of a word w in a document d to zero if the word appears in all documents, and increases it gradually as the number of documents containing w decreases [7].

$$\mathit{tfidf}_{d,w} = \frac{freq_{w,d}}{|d|} \cdot log \frac{|\mathcal{D}|}{|\{d \in \mathcal{D} : w \in d\}|} \tag{5}$$

where $|\cdot|$ is size of a set; $freq_{w,d}$ is the number of occurrences of word w in document d.

As the number of words in a large document collection could surpass 10^6 (which would result in up to 10^{12} possible bigrams, if these were used), dimension-reducing techniques can also be considered to reduce the dimensionality of the representation in order to fight increase in computational complexity and sparsity of data. One such approach that is quickly growing in popularity is *word2vec* [8], which uses the class of neural networks popularised under the label of Deep Learning to reduce the representation dimensionality to value k which is typically $100 < k < 1000$. The result is that each word is represented as a linear combination of these new features, that is, a vector of size k known as *word embedding*. We then represent a document of n words as the average of its n word embeddings. A set of m documents is then represented as a matrix of size $m \times k$. The method relies on distributional statistics of words within a fixed-size window. These are often collected from very large corpora and then used with other documents of interest.

In this study, we always preprocess all text documents in the following way. First, the text is *tokenized*, i.e. split into separate words or punctuation symbols. Then we remove all punctuation and *stop words*, essentially all pronouns, prepositions, conjunctions and a few very common verbs. The remaining words are *lemmatized*, i.e. replaced by their standard entry in the dictionary. All URLs are then mapped to the same string (URL), email addresses are mapped to the string EMAIL, and numbers are mapped to NUM. Finally, we merge all preprocessed news items for each company into a single document. From this, we produce two representations of all news on m companies making use of n different words. One is the TF-IDF weighted bag-of-words $m \times n$ matrix, the other – the $m \times k$ matrix produced with the word2vec approach (where $k = 300$). For each of these representations we have experimented with two different distance metrics, CD and ED (as defined in Section III-B) to produce four different distance matrices representing how the news about our 27 companies differ from each other.

D. Modelling Stock Prices

Our data set contains the daily $open$, $high$, low, $close$ prices for each company, as well and the $volume$ of trade on that day.

Here $close$ refers to the final price of last deal before the stock market closes and we are using $adjusted\ close$, which refers to the price that depicts the effects of corporation actions such as dividends and stock split. $high$, resp. low refers to the highest, resp. lowest price achieved during the day. $volume$ is total number of shares traded on that day. We have also calculated the $overnight\ return$ according to equation 6 and $change$ according to equation 7.

$$
\begin{aligned}
return_t &= (close_t - close_{t-1})/close_{t-1} && (6)\\
change_t &= (high_t - low_t)/open_t && (7)
\end{aligned}
$$

In this study we have used in turn data on $close$, $volume$, $return$ and $change$ to produce distance matrices in each case using each of the three distance metrics defined in Section III-B. In each sliding window, the $close$ and $volume$ time series were *standarlized* according to the equation $s' = (s_t - \bar{s})/\sigma_{\mathbf{s}}$. Given a window from t_{start} to t_{end}, each variable will be represented as a vector: $< s'_{t_{start}}, s'_{t_{start}+1}, \cdots, s'_{t_{end}} >$

TABLE I: Experimental Settings

Setting	Options
News Representation	$tfidf$, $word2vec$
Text Distance Metric	CD, ED
Time Series	$close$, $volume$, $return$, $change$
Time Series Distance Metric	CD, ED, PD

$\mathbf{u}_x$: bag-of-words representation of news about stock x (e.g. $AAPL$)
$\mathbf{u}_y$: bag-of-words representation of news about stock y (e.g. $GOOG$)
$\mathbf{v}_x$: time series of stock x
$\mathbf{v}_y$: time series of stock y
$\mathbf{U}$: distance matrix of news
$\mathbf{V}$: distance matrix of stock time series

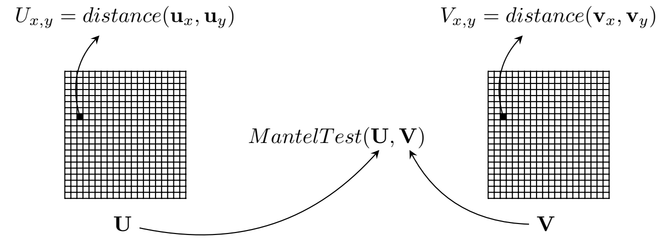

$U_{x,y} = distance(\mathbf{u}_x, \mathbf{u}_y)$
$V_{x,y} = distance(\mathbf{v}_x, \mathbf{v}_y)$
$MantelTest(\mathbf{U}, \mathbf{V})$
$\mathbf{U}$
$\mathbf{V}$

Fig. 1: Illustration of one Mantel test for an observation period.

IV. Experiment Design and Results

The first set of experiments considers all data available from 1 Oct 2014 until 30 April 2015. For each combination of time series, time series distance metric, text representation and text distance metric (see Table I), we performed the Mantel test to measure the correlation between news and time series. The results for all 48 combinations of experimental settings are reported in Table II along with the p-value of each test.

In our final set of experiments, we wanted to see whether the length of the time period used in the tests affected the levels of correlation, and to what extent the correlation varied over time. For that purpose, we used a 28-day long sliding window, and gradually shifted it with a 1-day step to produce 185 samples. All 48 experimental settings were then applied in turn to each sample, with the results plotted in the form of graphs, as shown in Figures 2 and 3.

V. Discussion

The results with the full data set suggest low to moderate levels of statistically significant correlation between news and financial performance, which for the best set of parameters reaches values of around 0.44. This, of course, does not indicate whether it is the news that affects the prices or, for instance, whether the news does not simply reflect the numerical data with a certain delay, which is likely to render it useless for forecasting. We do not attempt to answer this question here. On the other hand, the levels of correlation, achieved without any optimisation process in the choice of text-based features should serve as an encouragement to further studies, in which the most useful text features, be it words, bigrams, syntactic trees, etc. could be detected, either for the whole area of financial forecasting, or for a selected set of sectors.

The results show that text representations using $tfidf$ consistently outperform $word2vec$ and result in higher Mantel correlations between text and time series. $volume$ (*standarlized*) shows the highest correlation with news when $tfidf$ is used. Overnight $return$ also shows significant correlation with news of 0.35.

The results with the 28-day sliding window data show statistically significant results for extended periods of time with the correlation reaching levels of over 0.45 in some cases. There is a difference among the 4 variables representing stocks, with the *standardlized* $volume$ again showing the strongest correlation over the longest periods of time. Overnight returns also show substantial levels of correlation with news, albeit less often, and to a lower degree. It is also very interesting to observe sharp changes in the correlation levels, which could be potentially useful to detect important events as the first step towards forecasts that take into account the effect of such externalities.

It is worth mentioning that according to Augmented Dickey Fuller test, the $close$ price of all 27 stocks in our long observation time period is not stationary or trend stationary, thus PD is not a suitable distance metric, since the presence of trends in a pair of time series will boost the levels of correlation reported. We have kept these figures here for the sake of completeness.

$tfidf$ outperforms $word2vec$ with higher Mantel correlations, and, in the case of the sliding window data, also yields longer periods of statistical significance. Nevertheless, the plots also show that there are times when $word2vec$ is better than $tfidf$ at capturing significant correlations.

TABLE II: Correlation between news and financial time series as measured by the Mantel test according to subsection III-A ($10,000$ permutations for each Mantel test). A p-value for each Mantel test is also reported (in brackets); p-values less than 0.05 are shown in bold.

		$tfidf$		$word2vec$	
		CD	ED	CD	ED
$close$	CD	0.0820 (0.3947)	0.1581 (0.1210)	0.0108 (0.9006)	0.0196 (0.8061)
	ED	0.0922 (0.3413)	0.1578 (0.1364)	0.0253 (0.7797)	0.0373 (0.6558)
	PD*	0.0820 (0.3905)	0.1581 (0.1257)	0.0108 (0.9028)	0.0196 (0.8182)
$volume$	CD	**0.4214** **(0.0003)**	**0.4054** **(0.0012)**	0.0084 (0.9374)	0.0591 (0.5765)
	ED	**0.4433** **(0.0001)**	**0.4232** **(0.0006)**	0.0230 (0.8225)	0.0799 (0.4231)
	PD	**0.4214** **(0.0002)**	**0.4054** **(0.0016)**	0.0084 (0.9347)	0.0591 (0.5595)
$return$	CD	**0.3553** **(0.0013)**	**0.3476** **(0.0043)**	0.1695 (0.0855)	**0.1998** **(0.0388)**
	ED	0.2105 (0.1983)	0.2507 (0.1632)	−0.0459 (0.7331)	−0.0338 (0.7960)
	PD	**0.3567** **(0.0007)**	**0.3479** **(0.0051)**	0.1687 (0.0947)	**0.1991** **(0.0393)**
$change$	CD	0.1724 (0.2465)	0.1673 (0.3110)	0.2415 (0.0650)	0.1122 (0.3587)
	ED	0.2839 (0.0663)	**0.3555** **(0.0338)**	0.1190 (0.3654)	0.0149 (0.9093)
	PD	**0.3491** **(0.0036)**	**0.3341** **(0.0121)**	0.0484 (0.6669)	0.0961 (0.3618)

(* PD is not stationary)

References

[1] D. Tweney, "Twitter-fueled hedge fund bit the dust, but it actually worked," http://venturebeat.com/2012/05/28/twitter-fueled-hedge-fund-bit-the-dust-but-it-actually-worked/, 2012, [Online; accessed 12-Oct-2016].

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[3] H. Qu, M. Sardelich, N. N. Qomariyah, and D. Kazakov, "Integrating time series with social media data in an ontology for the modelling of extreme financial events," in *LREC 2016 Joint Second Workshop on Language and Ontology & Terminology and Knowledge Structures*, 2016.

[4] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer research*, vol. 27, no. 2.

[5] A. Moschitti, "Efficient convolution kernels for dependency and constituent syntactic trees," in *European Conference on Machine Learning*. Springer, 2006, pp. 318–329.

[6] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[7] J. Sedding and D. Kazakov, "Wordnet-based text document clustering," in *COLING 2004 3rd Workshop on Robust Methods in Analysis of Natural Language Data*, V. Pallotta and A. Todirascu, Eds., Geneva, 2004.

[8] T. Mikolov and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 2013.

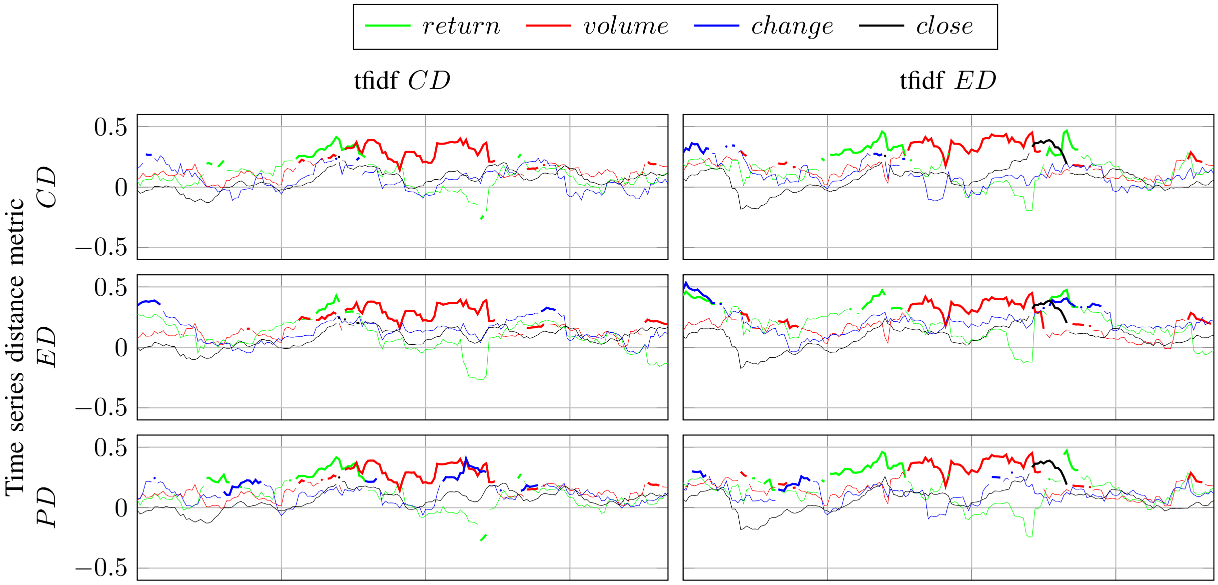


Fig. 2: Mantel test correlations for a sliding window of 4 weeks. Bold lines indicate statistically significant results ($p < 0.05$).

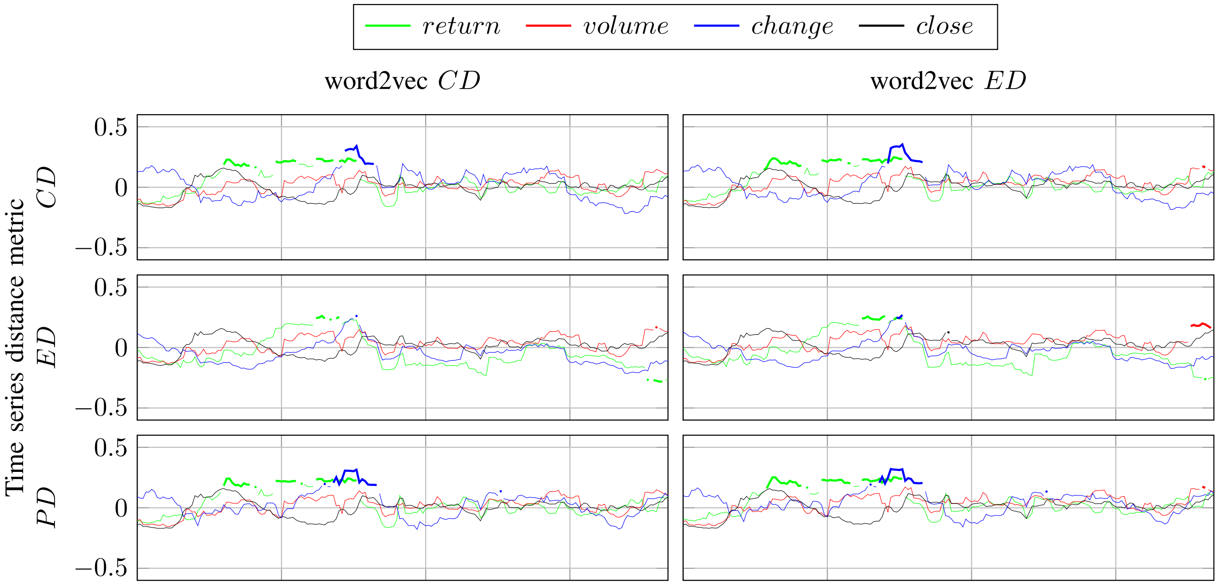
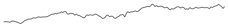
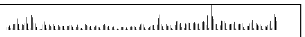
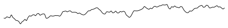
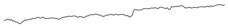
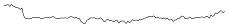
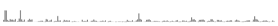
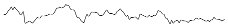
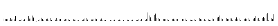


Fig. 3: Mantel test correlations for a sliding window of 4 weeks. Bold lines indicate statistically significant results ($p < 0.05$).

TABLE III: Information for each company: symbol, name, stock exchange, close price and number of news per day.

Symbol	Name	Exchange	Stock Price	Number of News
AAPL	Apple Inc.	NASDAQ		
AMZN	Amazon.com Inc.	NASDAQ		
BA	Boeing Co.	NYSE		
CMCSA	Comcast Co.	NASDAQ		
CSCO	Cisco Systems, Inc.	NASDAQ		
CVX	Chevron Co.	NYSE		
DIS	Walt Disney Co.	NYSE		
EBAY	eBay Inc.	NASDAQ		
FB	Facebook Common Stock	NASDAQ		
GE	General Electric Co.	LON		
GOOG	Alphabet Inc. Class C	NASDAQ		
GOOGL	Alphabet Inc. Class A	NASDAQ		
GS	Goldman Sachs Group Inc.	NYSE		
HD	Home Depot Inc.	NYSE		
IBM	IBM Common Stock	LON		
INTC	Intel Co.	NASDAQ		
JPM	JPMorgan Chase & Co.	NYSE		
KO	The Coca-Cola Co.	NYSE		
MSFT	Microsoft Co.	NASDAQ		
NFLX	Netflix, Inc.	NASDAQ		
NKE	Nike Inc.	NYSE		
SBUX	Starbucks Co.	NASDAQ		
T	AT&T Inc.	NYSE		
TSLA	Tesla Motors Inc.	NASDAQ		
VZ	Verizon Communications Inc.	NYSE		
WMT	Wal-Mart Stores, Inc.	NYSE		
YHOO	Yahoo! Inc.	NASDAQ		