**Sentiment Analysis**

*The aim of this research is to expand the findings of the paper “Sentence by Sentence Sentiment Analysis” using refined sentence selection algorithm. The new sentences set has been then passed through the same methodology used in previous work and the same result is obtained even using a fairly broader dataset.*

**Data**

Data are composed by the articles corpus and wheat daily price index series published by the International Grains Council.

The first dataset represent wheat prices. Wheat daily price index provided by the International Grains Council is composed by 4408 daily observations spanning from 05/01/2000 to 26/11/2016. However it has to be underlined that Saturdays and Sundays do not have any price observation.

The second dataset is composed by sentiment observations extracted from articles sentences. In particular, 140971 articles scraped from the web on 27/11/2016 compose the articles corpus. Since such articles deal with many different topics, it has been developed a python program, called *The Reading Machine*, which processes the articles. After processing them, by topic they are classified into homogeneous clusters. After articles classification using *The Reading Machine*, 15826 agriculture-related articles are selected. Using a list of wheat-related keywords generated by Michael Kao’s *The Reading Machine*, articles in which these words appear are selected and then, always using the same wheat-related keywords set, wheat-related sentences are extracted (duplicated sentences are dropped). All the sentences are then analyzed using an unmodified version of VADER sentiment analysis tool: the result is a 24740 sentiment matrix composed by sentences dates, article ID, same sentence text and compound sentiment score. It’s important to underline that sentences number is bigger than previous analysis because on one hand more words have been considered for both articles and sentences extraction, while on the other hand articles have been cleaned from special characters, therefore many previously dropped off ‘problematic’ articles have been reintroduced.

As in the previous paper noisy compound sentiment scores are then filtered using Kalman Filter generating 24740 *Filtered Sentiment* scores.

As final step these two datasets are merged in a unique dataset where each entry is composed by *unevenly spaced* but *contemporaneous* daily IGC Wheat Price Index and daily *Filtered Sentiment* observations. This final dataset is composed by 954 observations ranging from 01/01/2010 to 25/11/2016 and it’s used to study Wheat Price Index and *Filtered Sentiment* correlation. Being approximately 10 times the number of observations used in previous analysis, this result can be easily explored. Moreover from 2010 to 2016 are registered approximately 140 observations per each year.

**Results**

Given the bigger number of observations, *Filtered Sentiment* stationarity appears now evident: it seems that *Filtered Sentiment* floats around zero. Such stationarity has been tested using an Augmented Dickey-Fuller test which rejected the null of non-stationarity. On the other hand prices have been proven to be non-stationary as normal, therefore the two processes are integrated of different orders, making the use of correlation quite unuseful for sharp quantitative measurements.

What’s interesting however is that *Filtered Sentiment* and *Wheat Price Index* observations seems to mantain that inverse relationship already evidenced.



It’s also possible to underline a weak negative relation by analyzing *Filtered Sentiment* and daily Wheat price scatterplot.



**Conclusion**

This study shows a *Filtered Sentiment* and Wheat prices inverse relation new evidence based on a broader dataset that can be used for further applications. However it has to be underlined that no significant correlation between *Filtered Sentiment* and first differences have been found.

Next Reasearch Topics

* Use the Sentence by Sentence Sentiment Extraction Methodology on geo-tagged articles and local price series.
* Constrain sentences set to the ones that contain only future information.
* Analyze the new sentences set to identify helpful additions to VADER lexicon.