Stat. 653 Homework 4

Jiayi Liu

April 22, 2019

Upload one file to Blackboard.

Read: Chapter 5 Read: Chapter 15 in the Modern Data Science with R Problems:

# Run the R code from Chapter 4. 05-document-term-matrices.Rmd

library(knitr)  
opts\_chunk$set(message = FALSE, warning = FALSE, cache = TRUE)  
options(width = 100, dplyr.width = 150)  
library(ggplot2)  
library(methods)  
theme\_set(theme\_light())  
library(purrr)

knitr::include\_graphics("images/tidyflow-ch-5.png")

library(tm)  
  
data("AssociatedPress", package = "topicmodels")  
AssociatedPress

## <<DocumentTermMatrix (documents: 2246, terms: 10473)>>  
## Non-/sparse entries: 302031/23220327  
## Sparsity : 99%  
## Maximal term length: 18  
## Weighting : term frequency (tf)

terms <- Terms(AssociatedPress)  
head(terms)

## [1] "aaron" "abandon" "abandoned" "abandoning" "abbott" "abboud"

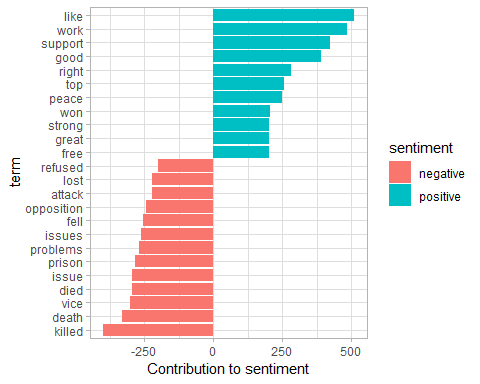
library(dplyr)  
library(tidytext)  
  
ap\_td <- tidy(AssociatedPress)  
ap\_td

## # A tibble: 302,031 x 3  
## document term count  
## <int> <chr> <dbl>  
## 1 1 adding 1  
## 2 1 adult 2  
## 3 1 ago 1  
## 4 1 alcohol 1  
## 5 1 allegedly 1  
## 6 1 allen 1  
## 7 1 apparently 2  
## 8 1 appeared 1  
## 9 1 arrested 1  
## 10 1 assault 1  
## # ... with 302,021 more rows

ap\_sentiments <- ap\_td %>%  
 inner\_join(get\_sentiments("bing"), by = c(term = "word"))  
  
ap\_sentiments

## # A tibble: 30,094 x 4  
## document term count sentiment  
## <int> <chr> <dbl> <chr>   
## 1 1 assault 1 negative   
## 2 1 complex 1 negative   
## 3 1 death 1 negative   
## 4 1 died 1 negative   
## 5 1 good 2 positive   
## 6 1 illness 1 negative   
## 7 1 killed 2 negative   
## 8 1 like 2 positive   
## 9 1 liked 1 positive   
## 10 1 miracle 1 positive   
## # ... with 30,084 more rows

library(ggplot2)  
  
ap\_sentiments %>%  
 count(sentiment, term, wt = count) %>%  
 ungroup() %>%  
 filter(n >= 200) %>%  
 mutate(n = ifelse(sentiment == "negative", -n, n)) %>%  
 mutate(term = reorder(term, n)) %>%  
 ggplot(aes(term, n, fill = sentiment)) +  
 geom\_bar(stat = "identity") +  
 ylab("Contribution to sentiment") +  
 coord\_flip()



### Tidying dfm objects

data("data\_corpus\_inaugural", package = "quanteda")  
inaug\_dfm <- quanteda::dfm(data\_corpus\_inaugural, verbose = FALSE)

inaug\_dfm

## Document-feature matrix of: 58 documents, 9,357 features (91.8% sparse).

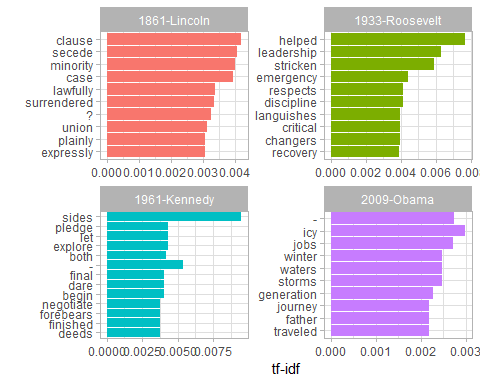
inaug\_td <- tidy(inaug\_dfm)  
inaug\_td

## # A tibble: 44,709 x 3  
## document term count  
## <chr> <chr> <dbl>  
## 1 1789-Washington fellow-citizens 1  
## 2 1797-Adams fellow-citizens 3  
## 3 1801-Jefferson fellow-citizens 2  
## 4 1809-Madison fellow-citizens 1  
## 5 1813-Madison fellow-citizens 1  
## 6 1817-Monroe fellow-citizens 5  
## 7 1821-Monroe fellow-citizens 1  
## 8 1841-Harrison fellow-citizens 11  
## 9 1845-Polk fellow-citizens 1  
## 10 1849-Taylor fellow-citizens 1  
## # ... with 44,699 more rows

inaug\_tf\_idf <- inaug\_td %>%  
 bind\_tf\_idf(term, document, count) %>%  
 arrange(desc(tf\_idf))  
  
inaug\_tf\_idf

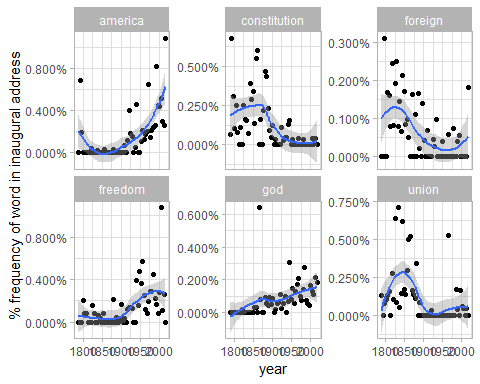
## # A tibble: 44,709 x 6  
## document term count tf idf tf\_idf  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 1793-Washington arrive 1 0.00680 4.06 0.0276  
## 2 1793-Washington upbraidings 1 0.00680 4.06 0.0276  
## 3 1793-Washington violated 1 0.00680 3.37 0.0229  
## 4 1793-Washington willingly 1 0.00680 3.37 0.0229  
## 5 1793-Washington incurring 1 0.00680 3.37 0.0229  
## 6 1793-Washington previous 1 0.00680 2.96 0.0201  
## 7 1793-Washington knowingly 1 0.00680 2.96 0.0201  
## 8 1793-Washington injunctions 1 0.00680 2.96 0.0201  
## 9 1793-Washington witnesses 1 0.00680 2.96 0.0201  
## 10 1793-Washington besides 1 0.00680 2.67 0.0182  
## # ... with 44,699 more rows

speeches <- c("1933-Roosevelt", "1861-Lincoln",  
 "1961-Kennedy", "2009-Obama")  
  
inaug\_tf\_idf %>%  
 filter(document %in% speeches) %>%  
 group\_by(document) %>%  
 top\_n(10, tf\_idf) %>%  
 ungroup() %>%  
 mutate(term = reorder(term, tf\_idf)) %>%  
 ggplot(aes(term, tf\_idf, fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~ document, scales = "free") +  
 coord\_flip() +  
 labs(x = "",  
 y = "tf-idf")



library(tidyr)  
  
year\_term\_counts <- inaug\_td %>%  
 extract(document, "year", "(\\d+)", convert = TRUE) %>%  
 complete(year, term, fill = list(count = 0)) %>%  
 group\_by(year) %>%  
 mutate(year\_total = sum(count))

year\_term\_counts %>%  
 filter(term %in% c("god", "america", "foreign", "union", "constitution", "freedom")) %>%  
 ggplot(aes(year, count / year\_total)) +  
 geom\_point() +  
 geom\_smooth() +  
 facet\_wrap(~ term, scales = "free\_y") +  
 scale\_y\_continuous(labels = scales::percent\_format()) +  
 ylab("% frequency of word in inaugural address")



## Casting tidy text data into a matrix

ap\_td %>%  
 cast\_dtm(document, term, count)

## <<DocumentTermMatrix (documents: 2246, terms: 10473)>>  
## Non-/sparse entries: 302031/23220327  
## Sparsity : 99%  
## Maximal term length: 18  
## Weighting : term frequency (tf)

ap\_td %>%  
 cast\_dfm(document, term, count)

## Document-feature matrix of: 2,246 documents, 10,473 features (98.7% sparse).

library(Matrix)  
  
# cast into a Matrix object  
m <- ap\_td %>%  
 cast\_sparse(document, term, count)  
  
class(m)

## [1] "dgCMatrix"  
## attr(,"package")  
## [1] "Matrix"

dim(m)

## [1] 2246 10473

library(janeaustenr)  
  
austen\_dtm <- austen\_books() %>%  
 unnest\_tokens(word, text) %>%  
 count(book, word) %>%  
 cast\_dtm(book, word, n)  
  
austen\_dtm

## <<DocumentTermMatrix (documents: 6, terms: 14520)>>  
## Non-/sparse entries: 40379/46741  
## Sparsity : 54%  
## Maximal term length: 19  
## Weighting : term frequency (tf)

## Tidying corpus objects with metadata

data("acq")  
acq

## <<VCorpus>>  
## Metadata: corpus specific: 0, document level (indexed): 0  
## Content: documents: 50

# first document  
acq[[1]]

## <<PlainTextDocument>>  
## Metadata: 15  
## Content: chars: 1287

acq\_td <- tidy(acq)  
acq\_td

## # A tibble: 50 x 16  
## author datetimestamp description heading id language origin topics lewissplit cgisplit  
## <chr> <dttm> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 <NA> 1987-02-26 07:18:06 "" COMPUTE~ 10 en Reute~ YES TRAIN TRAININ~  
## 2 <NA> 1987-02-26 07:19:15 "" OHIO MA~ 12 en Reute~ YES TRAIN TRAININ~  
## 3 <NA> 1987-02-26 07:49:56 "" MCLEAN'~ 44 en Reute~ YES TRAIN TRAININ~  
## 4 By Cal~ 1987-02-26 07:51:17 "" CHEMLAW~ 45 en Reute~ YES TRAIN TRAININ~  
## 5 <NA> 1987-02-26 08:08:33 "" <COFAB ~ 68 en Reute~ YES TRAIN TRAININ~  
## 6 <NA> 1987-02-26 08:32:37 "" INVESTM~ 96 en Reute~ YES TRAIN TRAININ~  
## 7 By Pat~ 1987-02-26 08:43:13 "" AMERICA~ 110 en Reute~ YES TRAIN TRAININ~  
## 8 <NA> 1987-02-26 08:59:25 "" HONG KO~ 125 en Reute~ YES TRAIN TRAININ~  
## 9 <NA> 1987-02-26 09:01:28 "" LIEBERT~ 128 en Reute~ YES TRAIN TRAININ~  
## 10 <NA> 1987-02-26 09:08:27 "" GULF AP~ 134 en Reute~ YES TRAIN TRAININ~  
## oldid places people orgs exchanges text   
## <chr> <list> <lgl> <lgl> <lgl> <chr>   
## 1 5553 <chr [1~ NA NA NA "Computer Terminal Systems Inc said\nit has completed the ~  
## 2 5555 <chr [1~ NA NA NA "Ohio Mattress Co said its first\nquarter, ending February~  
## 3 5587 <chr [1~ NA NA NA "McLean Industries Inc's United\nStates Lines Inc subsidia~  
## 4 5588 <chr [1~ NA NA NA "ChemLawn Corp <CHEM> could attract a\nhigher bid than the~  
## 5 5611 <chr [1~ NA NA NA "CoFAB Inc said it acquired <Gulfex Inc>,\na Houston-based~  
## 6 5639 <chr [1~ NA NA NA "A group of affiliated New York\ninvestment firms said the~  
## 7 5653 <chr [1~ NA NA NA "American Express Co remained silent on\nmarket rumors it ~  
## 8 5668 <chr [1~ NA NA NA "Industrial Equity (Pacific) Ltd, a\nHong Kong investment ~  
## 9 5671 <chr [1~ NA NA NA "Liebert Corp said its shareholders\napproved the merger o~  
## 10 5677 <chr [1~ NA NA NA "Gulf Applied Technologies Inc said it\nsold its subsidiar~  
## # ... with 40 more rows

acq\_tokens <- acq\_td %>%  
 select(-places) %>%  
 unnest\_tokens(word, text) %>%  
 anti\_join(stop\_words, by = "word")  
  
# most common words  
acq\_tokens %>%  
 count(word, sort = TRUE)

## # A tibble: 1,566 x 2  
## word n  
## <chr> <int>  
## 1 dlrs 100  
## 2 pct 70  
## 3 mln 65  
## 4 company 63  
## 5 shares 52  
## 6 reuter 50  
## 7 stock 46  
## 8 offer 34  
## 9 share 34  
## 10 american 28  
## # ... with 1,556 more rows

# tf-idf  
acq\_tokens %>%  
 count(id, word) %>%  
 bind\_tf\_idf(word, id, n) %>%  
 arrange(desc(tf\_idf))

## # A tibble: 2,853 x 6  
## id word n tf idf tf\_idf  
## <chr> <chr> <int> <dbl> <dbl> <dbl>  
## 1 186 groupe 2 0.133 3.91 0.522  
## 2 128 liebert 3 0.130 3.91 0.510  
## 3 474 esselte 5 0.109 3.91 0.425  
## 4 371 burdett 6 0.103 3.91 0.405  
## 5 442 hazleton 4 0.103 3.91 0.401  
## 6 199 circuit 5 0.102 3.91 0.399  
## 7 162 suffield 2 0.1 3.91 0.391  
## 8 498 west 3 0.1 3.91 0.391  
## 9 441 rmj 8 0.121 3.22 0.390  
## 10 467 nursery 3 0.0968 3.91 0.379  
## # ... with 2,843 more rows

### Example: mining financial articles

library(tm.plugin.webmining)  
library(purrr)  
  
library(tidyverse)  
library(mdsr)  
library(aRxiv)  
library(dplyr)

arxiv\_search(query = "MSFT")

## id submitted updated  
## 1 hep-th/0211238v3 2002-11-25 12:01:06 2004-05-22 18:45:14  
## 2 hep-th/0304005v3 2003-04-01 15:02:35 2003-07-20 19:37:18  
## 3 hep-th/0311264v2 2003-11-27 11:33:22 2003-12-09 15:42:45  
## 4 hep-th/0403283v1 2004-03-30 12:39:39 2004-03-30 12:39:39  
## 5 0807.3464v2 2008-07-22 12:50:31 2008-10-13 11:34:54  
## title  
## 1 MSFT : Moyal Star Formulation of String Field Theory  
## 2 Fermionic Ghosts in Moyal String Field Theory  
## 3 Improved Off-Shell Scattering Amplitudes in String Field Theory and New\n Computational Methods  
## 4 K-theory in cutoff version of Vacuum String Field Theory  
## 5 Joint analysis and estimation of stock prices and trading volume in\n Barndorff-Nielsen and Shephard stochastic volatility models  
## abstract  
## 1 The Moyal star formulation of string field theory is reviewed. The various\nversions of the star product are compared and related to one another in a\nregulated theory that resolves associativity anomalies. A summary of\ncomputations and challenges is given.\n  
## 2 We complete the construction of the Moyal star formulation of bosonic open\nstring field theory (MSFT) by providing a detailed study of the fermionic ghost\nsector. In particular, as in the case of the matter sector, (1) we construct a\nmap from Witten's star product to the Moyal product, (2) we propose a\nregularization scheme which is consistent with the matter sector and (3) as a\ncheck of the formalism, we derive the ghost Neumann coefficients algebraically\ndirectly from the Moyal product. The latter satisfy the Gross-Jevicki nonlinear\nrelations even in the presence of the regulator, and when the regulator is\nremoved they coincide numerically with the expression derived from conformal\nfield theory. After this basic construction, we derive a regularized action of\nstring field theory in the Siegel gauge and define the Feynman rules. We give\nexplicitly the analytic expression of the off-shell four point function for\ntachyons, including the ghost contribution. Some of the results in this paper\nhave already been used in our previous publications. This paper provides the\ntechnical details of the computations which were omitted there.\n  
## 3 We report on new results in Witten's cubic string field theory for the\noff-shell factor in the 4-tachyon amplitude that was not fully obtained\nexplicitly before. This is achieved by completing the derivation of the\nVeneziano formula in the Moyal star formulation of Witten's string field theory\n(MSFT). We also demonstrate detailed agreement of MSFT with a number of\non-shell and off-shell computations in other approaches to Witten's string\nfield theory. We extend the techniques of computation in MSFT, and show that\nthe j=0 representation of SL(2,R) generated by the Virasoro operators\n$L\_{0},L\_{\\pm1}$ is a key structure in practical computations for generating\nnumbers. We provide more insight into the Moyal structure that simplifies\nstring field theory, and develop techniques that could be applied more\ngenerally, including nonperturbative processes.\n  
## 4 Solutions of the Vacuum String Field Theory (VSFT) equation of motion\ninvolving matter part are given by projectors, and they represent\nnonperturbative solutions (e.g. the sliver) interpreted as D25-branes (or lower\ndimensional branes), but they are not mathematically well defined as they have\nzero norm. In this work we will use a regularization procedure based on the\ncutoff version of Moyal String Field Theory (MSFT), a particular version of\nVSFT, and we will see that both the sliver and the butterfly states, in this\nregime, have a good mathematical description. In particular they are\nexponential functions belonging to $\\Sc(\\RR^{2Nd})$, the space of Schwartzian\nfunctions equipped with the \*-product. Then we prove that if we classify those\nregularized solutions with K-theory group built out of the C\*-algebra\n$\\bar{\\Sc}(\\RR^{2Nd})$ we find exactly the same result obtained considering a\nK-theoretic classification of D25-branes in usual string theory, using the\ntopological K-theory of vector bundles over the D25-brane worldvolume. We then\ncomment on the meaning of this result and possible physical implications.\n  
## 5 We introduce a variant of the Barndorff-Nielsen and Shephard stochastic\nvolatility model where the non Gaussian Ornstein-Uhlenbeck process describes\nsome measure of trading intensity like trading volume or number of trades\ninstead of unobservable instantaneous variance. We develop an explicit\nestimator based on martingale estimating functions in a bivariate model that is\nnot a diffusion, but admits jumps. It is assumed that both the quantities are\nobserved on a discrete grid of fixed width, and the observation horizon tends\nto infinity. We show that the estimator is consistent and asymptotically normal\nand give explicit expressions of the asymptotic covariance matrix. Our method\nis illustrated by a finite sample experiment and a statistical analysis on the\nInternational Business Machines Corporation (IBM) stock from the New York Stock\nExchange (NYSE) and the Microsoft Corporation (MSFT) stock from Nasdaq during a\nhistory of five years.\n  
## authors affiliations link\_abstract  
## 1 Itzhak Bars http://arxiv.org/abs/hep-th/0211238v3  
## 2 I. Bars|I. Kishimoto|Y. Matsuo http://arxiv.org/abs/hep-th/0304005v3  
## 3 Itzhak Bars|I. Y. Park http://arxiv.org/abs/hep-th/0311264v2  
## 4 A. Parodi http://arxiv.org/abs/hep-th/0403283v1  
## 5 Friedrich Hubalek|Petra Posedel http://arxiv.org/abs/0807.3464v2  
## link\_pdf link\_doi  
## 1 http://arxiv.org/pdf/hep-th/0211238v3   
## 2 http://arxiv.org/pdf/hep-th/0304005v3 http://dx.doi.org/10.1088/1126-6708/2003/07/027  
## 3 http://arxiv.org/pdf/hep-th/0311264v2 http://dx.doi.org/10.1103/PhysRevD.69.086007  
## 4 http://arxiv.org/pdf/hep-th/0403283v1   
## 5 http://arxiv.org/pdf/0807.3464v2   
## comment journal\_ref doi  
## 1 22 pages, LaTeX. v3 minor corrections   
## 2 65 pages, typos corrected JHEP 0307 (2003) 027 10.1088/1126-6708/2003/07/027  
## 3 40 pages, 2 figures, LaTeX Phys.Rev. D69 (2004) 086007 10.1103/PhysRevD.69.086007  
## 4 24 pages, no figures   
## 5 26 pages, 16 figures   
## primary\_category categories  
## 1 hep-th hep-th  
## 2 hep-th hep-th  
## 3 hep-th hep-th  
## 4 hep-th hep-th  
## 5 q-fin.ST q-fin.ST|math.PR|math.ST|stat.TH|62H12; 60G42; 60J75

company <- c("Microsoft", "Apple", "Google", "Amazon", "Facebook",  
 "Twitter", "IBM", "Yahoo", "Netflix")  
symbol <- c("MSFT", "AAPL", "GOOG", "AMZN", "FB", "TWTR", "IBM", "YHOO", "NFLX")  
  
#download\_articles <- function(symbol) {  
 #p0=arxiv\_search(query = symbol)  
 # p0=paste0(p0)  
#}  
  
  
#stock\_articles <- data\_frame(company = company, symbol = symbol) %>%  
 # mutate(corpus = map(symbol, download\_articles))

p1 <- arxiv\_search(query = '"MSFT"', limit = 200)  
  
p1 <- p1 %>% mutate(company="Microsoft", symbol= "MSFT")  
head(p1)

## id submitted updated  
## 1 hep-th/0211238v3 2002-11-25 12:01:06 2004-05-22 18:45:14  
## 2 hep-th/0304005v3 2003-04-01 15:02:35 2003-07-20 19:37:18  
## 3 hep-th/0311264v2 2003-11-27 11:33:22 2003-12-09 15:42:45  
## 4 hep-th/0403283v1 2004-03-30 12:39:39 2004-03-30 12:39:39  
## 5 0807.3464v2 2008-07-22 12:50:31 2008-10-13 11:34:54  
## title  
## 1 MSFT : Moyal Star Formulation of String Field Theory  
## 2 Fermionic Ghosts in Moyal String Field Theory  
## 3 Improved Off-Shell Scattering Amplitudes in String Field Theory and New\n Computational Methods  
## 4 K-theory in cutoff version of Vacuum String Field Theory  
## 5 Joint analysis and estimation of stock prices and trading volume in\n Barndorff-Nielsen and Shephard stochastic volatility models  
## abstract  
## 1 The Moyal star formulation of string field theory is reviewed. The various\nversions of the star product are compared and related to one another in a\nregulated theory that resolves associativity anomalies. A summary of\ncomputations and challenges is given.\n  
## 2 We complete the construction of the Moyal star formulation of bosonic open\nstring field theory (MSFT) by providing a detailed study of the fermionic ghost\nsector. In particular, as in the case of the matter sector, (1) we construct a\nmap from Witten's star product to the Moyal product, (2) we propose a\nregularization scheme which is consistent with the matter sector and (3) as a\ncheck of the formalism, we derive the ghost Neumann coefficients algebraically\ndirectly from the Moyal product. The latter satisfy the Gross-Jevicki nonlinear\nrelations even in the presence of the regulator, and when the regulator is\nremoved they coincide numerically with the expression derived from conformal\nfield theory. After this basic construction, we derive a regularized action of\nstring field theory in the Siegel gauge and define the Feynman rules. We give\nexplicitly the analytic expression of the off-shell four point function for\ntachyons, including the ghost contribution. Some of the results in this paper\nhave already been used in our previous publications. This paper provides the\ntechnical details of the computations which were omitted there.\n  
## 3 We report on new results in Witten's cubic string field theory for the\noff-shell factor in the 4-tachyon amplitude that was not fully obtained\nexplicitly before. This is achieved by completing the derivation of the\nVeneziano formula in the Moyal star formulation of Witten's string field theory\n(MSFT). We also demonstrate detailed agreement of MSFT with a number of\non-shell and off-shell computations in other approaches to Witten's string\nfield theory. We extend the techniques of computation in MSFT, and show that\nthe j=0 representation of SL(2,R) generated by the Virasoro operators\n$L\_{0},L\_{\\pm1}$ is a key structure in practical computations for generating\nnumbers. We provide more insight into the Moyal structure that simplifies\nstring field theory, and develop techniques that could be applied more\ngenerally, including nonperturbative processes.\n  
## 4 Solutions of the Vacuum String Field Theory (VSFT) equation of motion\ninvolving matter part are given by projectors, and they represent\nnonperturbative solutions (e.g. the sliver) interpreted as D25-branes (or lower\ndimensional branes), but they are not mathematically well defined as they have\nzero norm. In this work we will use a regularization procedure based on the\ncutoff version of Moyal String Field Theory (MSFT), a particular version of\nVSFT, and we will see that both the sliver and the butterfly states, in this\nregime, have a good mathematical description. In particular they are\nexponential functions belonging to $\\Sc(\\RR^{2Nd})$, the space of Schwartzian\nfunctions equipped with the \*-product. Then we prove that if we classify those\nregularized solutions with K-theory group built out of the C\*-algebra\n$\\bar{\\Sc}(\\RR^{2Nd})$ we find exactly the same result obtained considering a\nK-theoretic classification of D25-branes in usual string theory, using the\ntopological K-theory of vector bundles over the D25-brane worldvolume. We then\ncomment on the meaning of this result and possible physical implications.\n  
## 5 We introduce a variant of the Barndorff-Nielsen and Shephard stochastic\nvolatility model where the non Gaussian Ornstein-Uhlenbeck process describes\nsome measure of trading intensity like trading volume or number of trades\ninstead of unobservable instantaneous variance. We develop an explicit\nestimator based on martingale estimating functions in a bivariate model that is\nnot a diffusion, but admits jumps. It is assumed that both the quantities are\nobserved on a discrete grid of fixed width, and the observation horizon tends\nto infinity. We show that the estimator is consistent and asymptotically normal\nand give explicit expressions of the asymptotic covariance matrix. Our method\nis illustrated by a finite sample experiment and a statistical analysis on the\nInternational Business Machines Corporation (IBM) stock from the New York Stock\nExchange (NYSE) and the Microsoft Corporation (MSFT) stock from Nasdaq during a\nhistory of five years.\n  
## authors affiliations link\_abstract  
## 1 Itzhak Bars http://arxiv.org/abs/hep-th/0211238v3  
## 2 I. Bars|I. Kishimoto|Y. Matsuo http://arxiv.org/abs/hep-th/0304005v3  
## 3 Itzhak Bars|I. Y. Park http://arxiv.org/abs/hep-th/0311264v2  
## 4 A. Parodi http://arxiv.org/abs/hep-th/0403283v1  
## 5 Friedrich Hubalek|Petra Posedel http://arxiv.org/abs/0807.3464v2  
## link\_pdf link\_doi  
## 1 http://arxiv.org/pdf/hep-th/0211238v3   
## 2 http://arxiv.org/pdf/hep-th/0304005v3 http://dx.doi.org/10.1088/1126-6708/2003/07/027  
## 3 http://arxiv.org/pdf/hep-th/0311264v2 http://dx.doi.org/10.1103/PhysRevD.69.086007  
## 4 http://arxiv.org/pdf/hep-th/0403283v1   
## 5 http://arxiv.org/pdf/0807.3464v2   
## comment journal\_ref doi  
## 1 22 pages, LaTeX. v3 minor corrections   
## 2 65 pages, typos corrected JHEP 0307 (2003) 027 10.1088/1126-6708/2003/07/027  
## 3 40 pages, 2 figures, LaTeX Phys.Rev. D69 (2004) 086007 10.1103/PhysRevD.69.086007  
## 4 24 pages, no figures   
## 5 26 pages, 16 figures   
## primary\_category categories company symbol  
## 1 hep-th hep-th Microsoft MSFT  
## 2 hep-th hep-th Microsoft MSFT  
## 3 hep-th hep-th Microsoft MSFT  
## 4 hep-th hep-th Microsoft MSFT  
## 5 q-fin.ST q-fin.ST|math.PR|math.ST|stat.TH|62H12; 60G42; 60J75 Microsoft MSFT

library(lubridate)   
  
p1 <- p1 %>%  
 mutate(submitted = ymd\_hms(submitted), updated = ymd\_hms(updated))   
glimpse(p1)

## Observations: 5  
## Variables: 17  
## $ id <chr> "hep-th/0211238v3", "hep-th/0304005v3", "hep-th/0311264v2", "hep-th/0403283v1", "0807.3464v2"  
## $ submitted <dttm> 2002-11-25 12:01:06, 2003-04-01 15:02:35, 2003-11-27 11:33:22, 2004-03-30 12:39:39, 2008-07-22 12:50:31  
## $ updated <dttm> 2004-05-22 18:45:14, 2003-07-20 19:37:18, 2003-12-09 15:42:45, 2004-03-30 12:39:39, 2008-10-13 11:34:54  
## $ title <chr> "MSFT : Moyal Star Formulation of String Field Theory", "Fermionic Ghosts in Moyal String Field Theory", "Improved Off-S...  
## $ abstract <chr> " The Moyal star formulation of string field theory is reviewed. The various\nversions of the star product are compared...  
## $ authors <chr> "Itzhak Bars", "I. Bars|I. Kishimoto|Y. Matsuo", "Itzhak Bars|I. Y. Park", "A. Parodi", "Friedrich Hubalek|Petra Posedel"  
## $ affiliations <chr> "", "", "", "", ""  
## $ link\_abstract <chr> "http://arxiv.org/abs/hep-th/0211238v3", "http://arxiv.org/abs/hep-th/0304005v3", "http://arxiv.org/abs/hep-th/0311264v2...  
## $ link\_pdf <chr> "http://arxiv.org/pdf/hep-th/0211238v3", "http://arxiv.org/pdf/hep-th/0304005v3", "http://arxiv.org/pdf/hep-th/0311264v2...  
## $ link\_doi <chr> "", "http://dx.doi.org/10.1088/1126-6708/2003/07/027", "http://dx.doi.org/10.1103/PhysRevD.69.086007", "", ""  
## $ comment <chr> "22 pages, LaTeX. v3 minor corrections", "65 pages, typos corrected", "40 pages, 2 figures, LaTeX", "24 pages, no figure...  
## $ journal\_ref <chr> "", "JHEP 0307 (2003) 027", "Phys.Rev. D69 (2004) 086007", "", ""  
## $ doi <chr> "", "10.1088/1126-6708/2003/07/027", "10.1103/PhysRevD.69.086007", "", ""  
## $ primary\_category <chr> "hep-th", "hep-th", "hep-th", "hep-th", "q-fin.ST"  
## $ categories <chr> "hep-th", "hep-th", "hep-th", "hep-th", "q-fin.ST|math.PR|math.ST|stat.TH|62H12; 60G42; 60J75"  
## $ company <chr> "Microsoft", "Microsoft", "Microsoft", "Microsoft", "Microsoft"  
## $ symbol <chr> "MSFT", "MSFT", "MSFT", "MSFT", "MSFT"

tally(~ year(submitted), data = p1)

## year(submitted)  
## 2002 2003 2004 2008   
## 1 2 1 1

p1 %>% filter(year(submitted) == 2003) %>%   
 glimpse()

## Observations: 2  
## Variables: 17  
## $ id <chr> "hep-th/0304005v3", "hep-th/0311264v2"  
## $ submitted <dttm> 2003-04-01 15:02:35, 2003-11-27 11:33:22  
## $ updated <dttm> 2003-07-20 19:37:18, 2003-12-09 15:42:45  
## $ title <chr> "Fermionic Ghosts in Moyal String Field Theory", "Improved Off-Shell Scattering Amplitudes in String Field Theory and Ne...  
## $ abstract <chr> " We complete the construction of the Moyal star formulation of bosonic open\nstring field theory (MSFT) by providing a...  
## $ authors <chr> "I. Bars|I. Kishimoto|Y. Matsuo", "Itzhak Bars|I. Y. Park"  
## $ affiliations <chr> "", ""  
## $ link\_abstract <chr> "http://arxiv.org/abs/hep-th/0304005v3", "http://arxiv.org/abs/hep-th/0311264v2"  
## $ link\_pdf <chr> "http://arxiv.org/pdf/hep-th/0304005v3", "http://arxiv.org/pdf/hep-th/0311264v2"  
## $ link\_doi <chr> "http://dx.doi.org/10.1088/1126-6708/2003/07/027", "http://dx.doi.org/10.1103/PhysRevD.69.086007"  
## $ comment <chr> "65 pages, typos corrected", "40 pages, 2 figures, LaTeX"  
## $ journal\_ref <chr> "JHEP 0307 (2003) 027", "Phys.Rev. D69 (2004) 086007"  
## $ doi <chr> "10.1088/1126-6708/2003/07/027", "10.1103/PhysRevD.69.086007"  
## $ primary\_category <chr> "hep-th", "hep-th"  
## $ categories <chr> "hep-th", "hep-th"  
## $ company <chr> "Microsoft", "Microsoft"  
## $ symbol <chr> "MSFT", "MSFT"

tally(~ primary\_category, data = p1)

## primary\_category  
## hep-th q-fin.ST   
## 4 1

p1 %>% mutate(field = str\_extract(primary\_category, "^[a-z,-]+")) %>%   
 tally(x = ~field) %>%   
 sort()

## field  
## q-fin hep-th   
## 1 4

library(tm)   
  
Corpus <- with(p1, VCorpus(VectorSource(abstract)))   
Corpus[[1]] %>% as.character() %>%   
 strwrap()

## [1] "The Moyal star formulation of string field theory is reviewed. The various versions of"  
## [2] "the star product are compared and related to one another in a regulated theory that"   
## [3] "resolves associativity anomalies. A summary of computations and challenges is given."

Corpus <- Corpus %>% tm\_map(stripWhitespace) %>%   
 tm\_map(removeNumbers) %>%   
 tm\_map(removePunctuation) %>%   
 tm\_map(content\_transformer(tolower)) %>%   
 tm\_map(removeWords, stopwords("english"))  
strwrap(as.character(Corpus[[1]]))

## [1] "moyal star formulation string field theory reviewed various versions star product"   
## [2] "compared related one another regulated theory resolves associativity anomalies summary"  
## [3] "computations challenges given"

library(wordcloud)   
  
wordcloud(Corpus, max.words = 30, scale = c(8, 1), colors = topo.colors(n = 30), random.color = TRUE)



DTM <- DocumentTermMatrix(Corpus, control = list(weighting = weightTfIdf))   
DTM

## <<DocumentTermMatrix (documents: 5, terms: 268)>>  
## Non-/sparse entries: 312/1028  
## Sparsity : 77%  
## Maximal term length: 17  
## Weighting : term frequency - inverse document frequency (normalized) (tf-idf)

findFreqTerms(DTM, lowfreq = 0.8)

## character(0)

DTM %>% as.matrix() %>%   
 apply(MARGIN = 2, sum) %>%   
 sort(decreasing = TRUE) %>%   
 head(9)

## anomalies another associativity challenges compared one regulated   
## 0.096747 0.096747 0.096747 0.096747 0.096747 0.096747 0.096747   
## related resolves   
## 0.096747 0.096747

findAssocs(DTM, terms = "statistics", corlimit = 0.5)

## $statistics  
## numeric(0)

findAssocs(DTM, terms = "mathematics", corlimit = 0.5)

## $mathematics  
## numeric(0)