Corpus Analysis from a Mathematical Perspective

Corpus Statistics Research Group launch event
Birmingham, 11th Feb 2016

Simon Preston (University of Nottingham)

Joint work with R. Carrington, A. Hennessey, M. Mahlberg, K. Severn, Y. van Gennip, V. Wiegand

Corpus as a mathematical object

THE POSTHUMOUS PAPERS
OF
THE PICKWICK CLUB

CHAPTER I
THE PICKWICKIANS

The first ray of light which illumines the gloom, and converts into a dazzling brilliancy that obscurity in which the earlier history of the public career of the immortal Pickwick would appear to be involved, is derived from the perusal of the following entry in the Transactions of the Pickwick Club, which the editor of these papers feels the highest pleasure in laying before his readers, as a proof of the careful attention, indefatigable assiduity, and nice discrimination, with which his search among the multifarious documents confided to him has been conducted.

'May 12, 1827. Joseph Smiggers, Esq., P.V.P.M.P.C. [Perpetual Vice-President--Member Pickwick Club], presiding. The following resolutions unanimously agreed to:--

'That this Association has heard read, with feelings of unmingled satisfaction, and unqualified approval, the paper communicated by Samuel Pickwick, Esq., G.C.M.P.C. [General Chairman--Member Pickwick Club], entitled "Speculations on the Source of the Hampstead Ponds, with some Observations on the Theory of Tittlebats;" and that this Association does hereby return its warmest thanks to the said Samuel Pickwick, Esq., G.C.M.P.C., for the same.

'That while this Association is deeply sensible of the advantages which must accrue to the cause of science, from the production to which they have just adverted--no less than from the unwearied researches of Samuel Pickwick, Esq., G.C.M.P.C., in Hornsey, Highgate, Brixton, and Camberwell--they cannot but entertain

OLIVER TWIST
OR
THE PARISH BOY'S PROGRESS
BY
CHARLES DICKENS

CHAPTER I

TREATS OF THE PLACE WHERE OLIVER TWIST WAS BOR
CIRCUMSTANCES ATTENDING HIS BIRTH

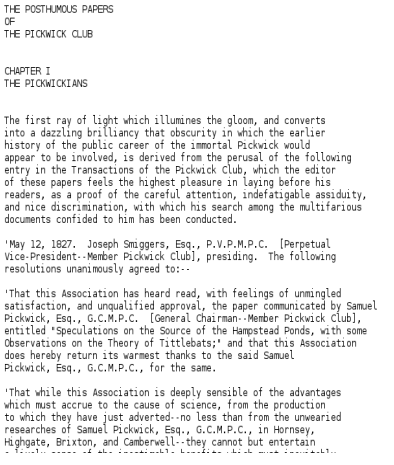
Among other public buildings in a certain town
reasons it will be prudent to refrain from men
which I will assign no fictitious name, there
common to most towns, great or small: to wit,
in this workhouse was born; on a day and date
trouble myself to repeat, inasmuch as it can b
consequence to the reader, in this stage of th
events; the item of mortality whose name is pr
of this chapter.

For a long time after it was ushered into this
and trouble, by the parish surgeon, it remaine
considerable doubt whether the child would sur
name at all; in which case it is somewhat more
these memoirs would never have appeared; or, i
being comprised within a couple of pages, they
possessed the inestimable merit of being the m
faithful specimen of biography, extant in the
age or country.

Although I am not disposed to maintain that th
workhouse, is in itself the most fortunate and
circumstance that can possibly befall a human
say that in this particular instance, it was t
Oliver Twist that could by possibility have oc

Corpus analysis

Corpus


THE POSTHUMOUS PAPERS
OF
THE PICKWICK CLUB

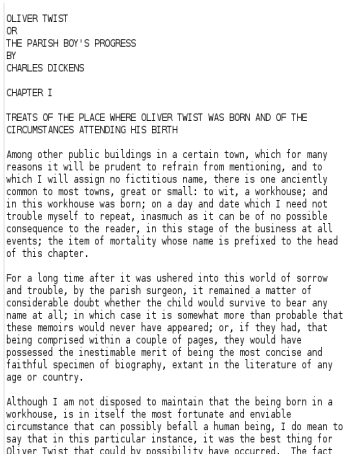
CHAPTER I
THE PICKWICKIANS

The first ray of light which illumines the gloom, and converts into a dazzling brilliancy that obscurity in which the earlier history of the public career of the immortal Pickwick would appear to be involved, is derived from the perusal of the following entry in the Transactions of the Pickwick Club, which the editor of these papers feels the highest pleasure in laying before his readers, as a proof of the careful attention, indefatigable assiduity, and nice discrimination, with which his search among the multifarious documents confided to him has been conducted.

'May 12, 1827. Joseph Smiggers, Esq., P.V.P.M.P.C. [Perpetual Vice-President--Member Pickwick Club], presiding. The following resolutions unanimously agreed to:--

'That this Association has heard read, with feelings of unmingled satisfaction, and unqualified approval, the paper communicated by Samuel Pickwick, Esq., G.C.M.P.C. [General Chairman--Member Pickwick Club], entitled 'Speculations on the Source of the Hampstead Ponds, with some Observations on the Theory of Tittlebats;' and that this Association does hereby return its warmest thanks to the said Samuel Pickwick, Esq., G.C.M.P.C., for the same.

'That while this Association is deeply sensible of the advantages which must accrue to the cause of science, from the production to which they have just adverted--no less than from the unwearied researches of Samuel Pickwick, Esq., G.C.M.P.C., in Hornsey, Highgate, Brixton, and Camberwell--they cannot but entertain


OLIVER TWIST
OR
THE PARISH BOY'S PROGRESS
BY
CHARLES DICKENS

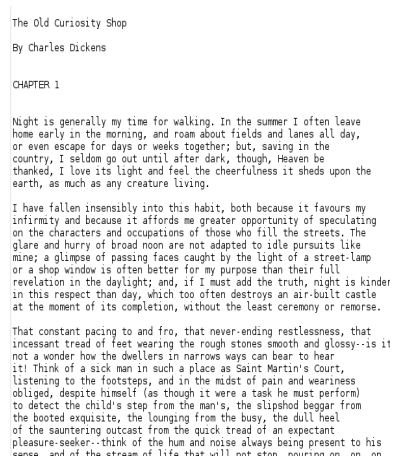
CHAPTER I

TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE CIRCUMSTANCES ATTENDING HIS BIRTH

Among other public buildings in a certain town, which for many reasons it will be prudent to refrain from mentioning, and to which I will assign no fictitious name, there is one anciently common to most towns, great or small: to wit, a workhouse; and in this workhouse was born; on a day and date which I need not trouble myself to repeat, inasmuch as it can be of no possible consequence to the reader, in this stage of the business at all events; the item of mortality whose name is prefixed to the head of this chapter.

For a long time after it was ushered into this world of sorrow and trouble, by the parish surgeon, it remained a matter of considerable doubt whether the child would survive to bear any name at all; in which case it is somewhat more than probable that these memoirs would never have appeared; or, if they had, that being comprised within a couple of pages, they would have possessed the inestimable merit of being the most concise and faithful specimen of biography, extant in the literature of any age or country.

Although I am not disposed to maintain that the being born in a workhouse, is in itself the most fortunate and enviable circumstance that can possibly befall a human being, I do mean to say that in this particular instance, it was the best thing for Oliver Twist that could by possibility have occurred. The fact


The Old Curiosity Shop

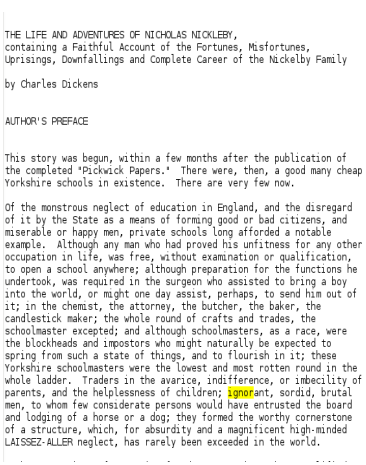
By Charles Dickens

CHAPTER 1

Night is generally my time for walking. In the summer I often leave home early in the morning, and roam about fields and lanes all day, or even escape for days or weeks together; but, saving in the country, I seldom go out until after dark, though, Heaven be thanked, I love its light and feel the cheerfulness it sheds upon the earth, as much as any creature living.

I have fallen insensibly into this habit, both because it favours my infirmity and because it affords me greater opportunity of speculating on the characters and occupations of those who fill the streets. The glare and hurry of broad noon are not adapted to idle pursuits like mine; a glimpse of passing faces caught by the light of a street-lamp or a shop window is often better for my purpose than their full revelation in the daylight; and, if I must add the truth, night is kinder in this respect than day, which too often destroys an air-built castle at the moment of its completion, without the least ceremony or remorse.

That constant pacing to and fro, that never-ending restlessness, that incessant tread of feet wearing the rough stones smooth and glossy--is it not a wonder how the dwellers in narrows ways can bear to hear it! Think of a sick man in such a place as Saint Martin's Court, listening to the footsteps, and in the midst of pain and weariness obliged, despite himself (as though it were a task he must perform) to detect the child's step from the man's, the slipshod beggar from the booted exquisite, the lounging from the busy, the dull heel of the sauntering outcast from the quick tread of an expectant pleasure-seeker--think of the hum and noise always being present to his


THE LIFE AND ADVENTURES OF NICHOLAS NICKLEBY,
containing a Faithful Account of the Fortunes, Misfortunes,
Uprisings, Downfallings and Complete Career of the Nickelby Family

by Charles Dickens

AUTHOR'S PREFACE

This story was begun, within a few months after the publication of the completed 'Pickwick Papers.' There were, then, a good many cheap Yorkshire schools in existence. There are very few now.

Of the monstrous neglect of education in England, and the disregard of it by the State as a means of forming good or bad citizens, and miserable or happy men, private schools long afforded a notable example. Although any man who had proved his unfitness for any other occupation in life, was free, without examination or qualification, to open a school anywhere; although preparation for the functions he undertook, was required in the surgeon who assisted to bring a boy into the world, or might one day assist, perhaps, to send him out of it; in the chemist, the attorney, the butcher, the baker, the candlestick maker; the whole round of crafts and trades, the schoolmaster excepted; and although schoolmasters, as a race, were the blockheads and impostors who might naturally be expected to spring from such a state of things, and to flourish in it; these Yorkshire schoolmasters were the lowest and most rotten round in the whole ladder. Traders in the avarice, indifference, or imbecility of parents, and the helplessness of children; ignorant, sordid, brutal men, to whom few considerate persons would have entrusted the board and lodging of a horse or a dog; they formed the worthy cornerstone of a structure, which, for absurdity and a magnificent high-minded LAISSEZ-ALLER neglect, has rarely been exceeded in the world.

Mathematical representation → X → Analysis → f(X)

Corpus analysis

Corpus

THE POSTHUMOUS PAPERS
OF
THE PICKWICK CLUB

CHAPTER I
THE PICKWICKIANS

The first ray of light which illumines the gloom, and converts into a dazzling brilliancy that obscurity in which the earlier history of the public career of the immortal Pickwick would appear to be involved, is derived from the perusal of the following entry in the Transactions of the Pickwick Club, which the editor of these papers feels the highest pleasure in laying before his readers, as a proof of the careful attention, indefatigable assiduity, and nice discrimination, with which his search among the multifarious documents confided to him has been conducted.

'May 12, 1827. Joseph Smiggers, Esq., P.V.P.M.P.C. [Perpetual Vice-President--Member Pickwick Club], presiding. The following resolutions unanimously agreed to:--

'That this Association has heard read, with feelings of unmingled satisfaction, and unqualified approval, the paper communicated by Samuel Pickwick, Esq., G.C.M.P.C. [General Chairman--Member Pickwick Club], entitled 'Speculations on the Source of the Hampstead Ponds, with some Observations on the Theory of Tittlebats;' and that this Association does hereby return its warmest thanks to the said Samuel Pickwick, Esq., G.C.M.P.C., for the same.

'That while this Association is deeply sensible of the advantages which must accrue to the cause of science, from the production to which they have just adverted--no less than from the unwearied researches of Samuel Pickwick, Esq., G.C.M.P.C., in Hornsey, Highgate, Brixton, and Camberwell--they cannot but entertain

OLIVER TWIST
OR
THE PARISH BOY'S PROGRESS
BY
CHARLES DICKENS

CHAPTER I

TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE CIRCUMSTANCES ATTENDING HIS BIRTH

Among other public buildings in a certain town, which for many reasons it will be prudent to refrain from mentioning, and to which I will assign no fictitious name, there is one anciently common to most towns, great or small: to wit, a workhouse; and in this workhouse was born; on a day and date which I need not trouble myself to repeat, inasmuch as it can be of no possible consequence to the reader, in this stage of the business at all events; the item of mortality whose name is prefixed to the head of this chapter.

For a long time after it was ushered into this world of sorrow and trouble, by the parish surgeon, it remained a matter of considerable doubt whether the child would survive to bear any name at all; in which case it is somewhat more than probable that these memoirs would never have appeared; or, if they had, that being comprised within a couple of pages, they would have possessed the inestimable merit of being the most concise and faithful specimen of biography, extant in the literature of any age or country.

Although I am not disposed to maintain that the being born in a workhouse, is in itself the most fortunate and enviable circumstance that can possibly befall a human being, I do mean to say that in this particular instance, it was the best thing for Oliver Twist that could by possibility have occurred. The fact

The Old Curiosity Shop

By Charles Dickens

CHAPTER 1

Night is generally my time for walking. In the summer I often leave home early in the morning, and roam about fields and lanes all day, or even escape for days or weeks together; but, saving in the country, I seldom go out until after dark, though, Heaven be thanked, I love its light and feel the cheerfulness it sheds upon the earth, as much as any creature living.

I have fallen insensibly into this habit, both because it favours my infirmity and because it affords me greater opportunity of speculating on the characters and occupations of those who fill the streets. The glare and hurry of broad noon are not adapted to idle pursuits like mine; a glimpse of passing faces caught by the light of a street-lamp or a shop window is often better for my purpose than their full revelation in the daylight; and, if I must add the truth, night is kinder in this respect than day, which too often destroys an air-built castle at the moment of its completion, without the least ceremony or remorse.

That constant pacing to and fro, that never-ending restlessness, that incessant tread of feet wearing the rough stones smooth and glossy--is it not a wonder how the dwellers in narrows ways can bear to hear it! Think of a sick man in such a place as Saint Martin's Court, listening to the footsteps, and in the midst of pain and weariness obliged, despite himself (as though it were a task he must perform) to detect the child's step from the man's, the slipshod beggar from the booted exquisite, the lounging from the busy, the dull heel of the sauntering outcast from the quick tread of an expectant pleasure-seeker--think of the hum and noise always being present to his

THE LIFE AND ADVENTURES OF NICHOLAS NICKLEBY,
containing a Faithful Account of the Fortunes, Misfortunes, Uprisings, Downfallings and Complete Career of the Nickelby Family

by Charles Dickens

AUTHOR'S PREFACE

This story was begun, within a few months after the publication of the completed 'Pickwick Papers.' There were, then, a good many cheap Yorkshire schools in existence. There are very few now.

Of the monstrous neglect of education in England, and the disregard of it by the State as a means of forming good or bad citizens, and miserable or happy men, private schools long afforded a notable example. Although any man who had proved his unfitness for any other occupation in life, was free, without examination or qualification, to open a school anywhere; although preparation for the functions he undertook, was required in the surgeon who assisted to bring a boy into the world, or might one day assist, perhaps, to send him out of it; in the chemist, the attorney, the butcher, the baker, the candlestick maker; the whole round of crafts and trades, the schoolmaster excepted; and although schoolmasters, as a race, were the blockheads and impostors who might naturally be expected to spring from such a state of things, and to flourish in it; these Yorkshire schoolmasters were the lowest and most rotten round in the whole ladder. Traders in the avarice, indifference, or imbecility of parents, and the helplessness of children; ignorant, sordid, brutal men, to whom few considerate persons would have entrusted the board and lodging of a horse or a dog; they formed the worthy cornerstone of a structure, which, for absurdity and a magnificent high-minded LAISSEZ-ALLER neglect, has rarely been exceeded in the world.

Mathematical representation → X → f(X) Analysis

Analysis = studying patterns

- checking one is really there
- identifying new ones

Why is this perspective helpful?

- Deciding on X forces us to decide:
  - what in the corpus is important
  - what we are happy to discard
- For a given X we have a "toolbox" of available methods from which to choose f(X): the abstraction is powerful.
- It helps us understand the f(X) we choose to use.
- . . . which is essential for developing new methodologies.

Example: Dickens novels

X as "bag of words" representation.

	said	one	will	now	little	poor	upon	mrs	
PP	3321	766	437	471	651	95	608	508	
OT	1232	457	302	280	276	97	477	264	
NN	2706	1019	712	608	743	262	1065	1040	...
OCS	1420	653	331	436	646	177	796	252	
BR	1454	839	401	509	391	136	911	189	
MC	2786	1042	629	705	686	150	1153	953	
DS	2561	921	578	713	943	199	1105	1333	
DC	2950	908	531	741	1096	187	806	673	...
BH	1743	971	805	909	1152	230	786	677	
HT	727	292	233	268	200	58	285	392	
LD	2139	1000	663	661	1454	261	779	928	
TTC	661	438	290	262	267	87	289	18	
GE	1349	502	174	453	371	77	366	164	...
OMF	2180	859	622	757	878	252	753	988	
MED	406	229	266	206	203	70	227	77	

Such a "data matrix" is the central object in statistical multivariate analysis.

Analysis method: matrix factorisation

- Break down X into the product "A times B":

$$\underset{X}{\overset{\text{novel}\times\text{word}}{}} \approx \underset{A}{\overset{novel\times r}{}} \times \underset{B}{\overset{r\times word}{}}$$

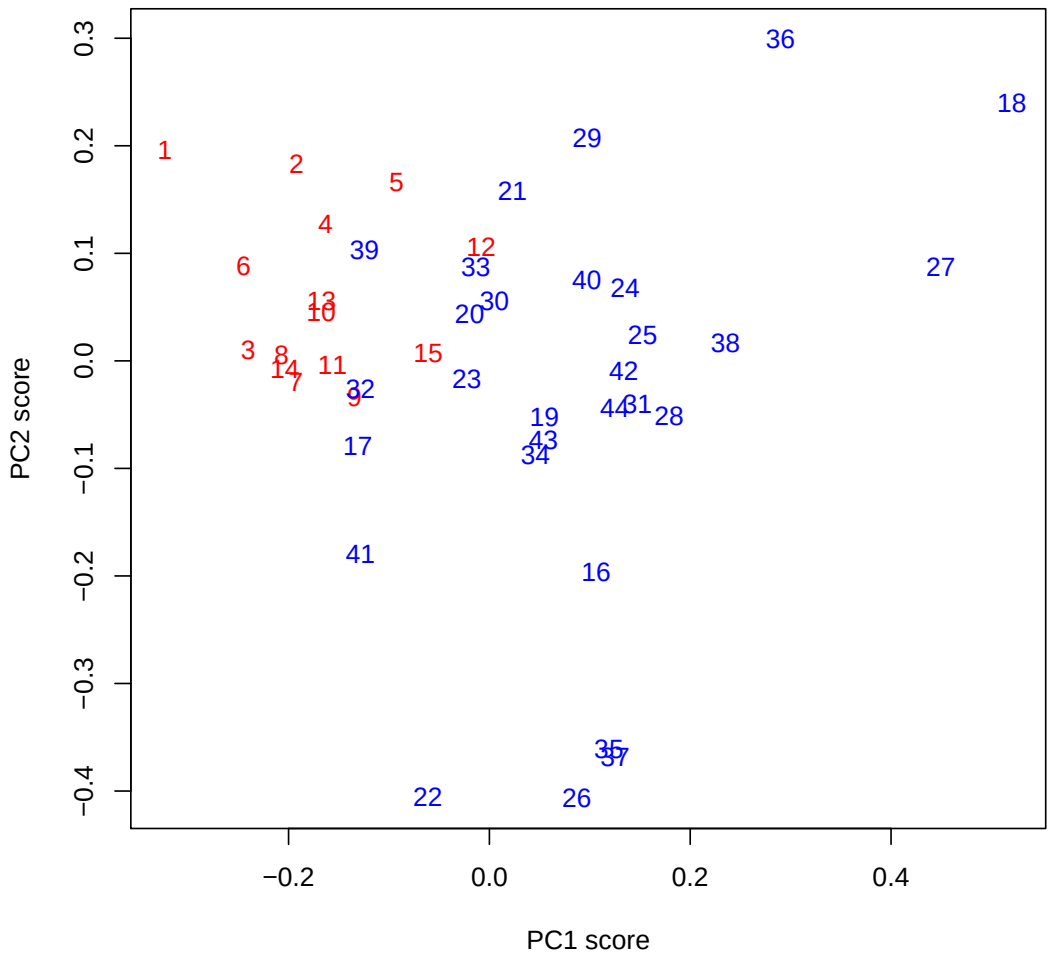
Analysis method: matrix factorisation

- Break down X into the product "A times B":

$$\underset{X}{\overset{\text{novel}\times\text{word}}{}} \approx \underset{A}{\overset{novel\times r}{}} \times \underset{B}{\overset{r\times word}{}}$$

- Rows of B represent "features" found in corpus.
- Rows of A represent novels as "scores" for these features.
- Different constraints on A and B results in well-known methods:
  - Principal component analysis (PCA)
  - Latent semantic analysis
  - Non-negative matrix factorisation
  - (Topic modelling)

PCA for Dickens and other 19C novels



Red = Dickens novels (numbering indicates chronology)

Blue = Misc other 19C novels (numbering arbitrary)

PC interpretation

- Interpretation of scores in A?
- First and second rows/features of B:

Row 1		Row 2	
said	-0.559	miss	-0.424
mrs	-0.184	mrs	-0.274
sir	-0.175	much	-0.129
old	-0.131	must	-0.127
upon	-0.125	little	-0.112
⋮	⋮	⋮	⋮
yet	0.128	man	0.122
will	0.143	upon	0.193
now	0.146	said	0.256

Other representations?

"Citizen Evremonde," she said, touching him with her cold hand. "I am a poor little seamstress, who was with you in La Force."

He murmured for answer: "True. I forget what you were accused of?"

"Plots. Though the just Heaven knows that I am innocent of any. Is it likely? Who would think of plotting with a poor little weak creature like me?"

The forlorn smile with which she said it, so touched him, that tears started from his eyes.

"I am not afraid to die, Citizen Evremonde, but I have done nothing. I am not unwilling to die, if the Republic which is to do so much good to us poor, will profit by my death; but I do not know how that can be, Citizen Evremonde. Such a poor weak little creature!"

As the last thing on earth that his heart was to warm and soften to, it warmed and softened to this pitiable girl.

"I heard you were released, Citizen Evremonde. I hoped it was true?"

"It was. But, I was again taken and condemned."

"If I may ride with you, Citizen Evremonde, will you let me hold your hand? I am not afraid, but I am little and weak, and it will give me more courage."

(A Tale of Two Cities, Dickens)

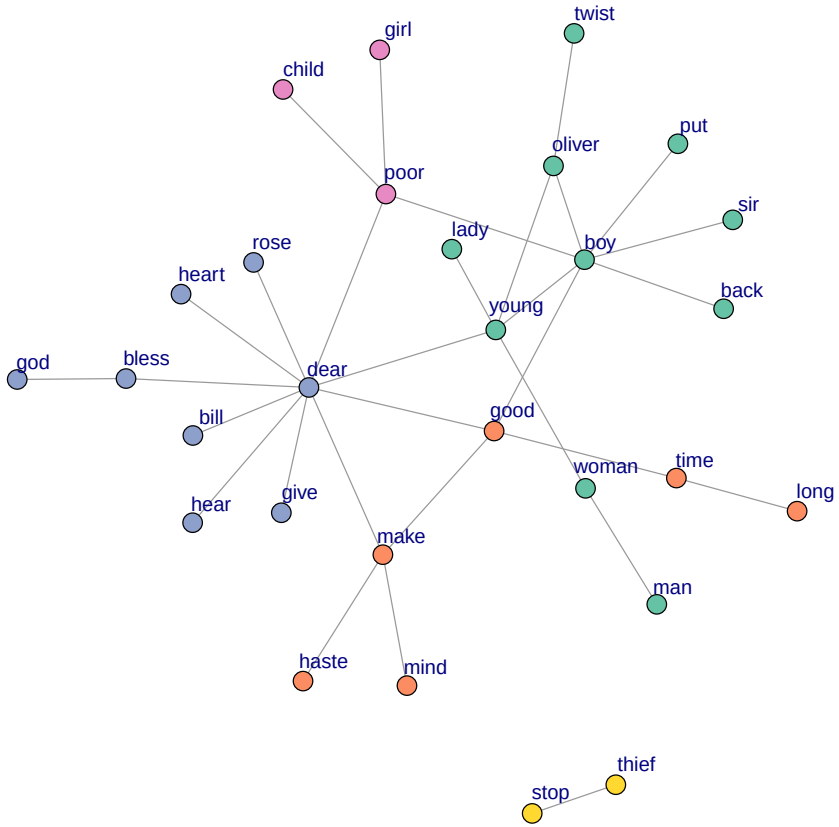
Speech from Oliver Twist: co-occurrence matrix

$$\underset{\text{word}\times\text{word}}{X} =$$

	dear	boy	good	bill	hear	sir	give	lady	haste	girl	bless	mind	oliver	stop	young	back	make	child	long	man	woman	time	heart	poor	god	put	twist	rose	thief
dear	22	5	8	9	9	4	7	5	0	4	8	3	3	3	10	3	7	6	2	2	1	1	10	7	5	0	1	7	0
boy	5	20	10	1	2	8	3	0	0	2	0	5	9	0	7	7	2	3	1	4	1	0	0	10	0	8	3	0	0
good	8	10	16	1	6	3	2	0	0	1	2	2	0	0	6	1	7	1	2	2	1	11	0	2	0	0	1	1	0
bill	9	1	1	12	1	0	3	0	0	0	0	2	0	0	1	0	0	1	1	1	0	3	1	2	2	0	0	0	0
hear	9	2	6	1	12	1	1	2	0	0	2	4	0	0	2	0	3	0	1	3	2	1	1	2	0	1	0	1	0
sir	4	8	3	0	1	12	0	0	0	0	0	0	4	1	1	0	0	1	1	1	0	2	1	3	0	0	2	0	0
give	7	3	2	3	1	0	10	3	1	1	0	2	2	1	2	3	2	0	2	3	0	3	1	1	1	2	0	0	0
lady	5	0	0	0	2	0	3	2	0	1	3	1	1	1	10	2	2	0	0	0	0	1	0	0	4	0	0	0	0
haste	0	0	0	0	0	0	1	0	2	0	0	2	0	0	0	0	9	0	0	1	0	0	0	0	0	0	0	0	0
girl	4	2	1	0	0	0	1	1	0	2	0	4	2	0	2	3	2	2	3	0	0	0	2	9	1	2	0	1	0
bless	8	0	2	0	2	0	0	3	0	0	2	1	0	0	0	1	0	0	1	1	0	0	6	2	8	0	0	0	0
mind	3	5	2	2	4	0	2	1	2	4	1	8	0	0	2	1	7	1	1	0	0	1	2	0	3	2	0	0	0
oliver	3	9	0	0	0	4	2	1	0	2	0	0	8	0	8	3	4	2	0	1	0	3	0	2	0	2	8	0	1
stop	3	0	0	0	0	1	1	1	0	0	0	0	0	8	0	0	0	0	1	1	0	0	0	0	1	0	0	0	7
young	10	7	6	1	2	1	2	10	0	2	0	2	8	0	8	0	2	2	1	5	7	5	4	6	0	4	4	0	0
back	3	7	1	0	0	0	3	2	0	3	1	1	3	0	0	6	3	0	4	6	0	5	1	2	3	1	0	0	2
make	7	2	7	0	3	0	2	2	9	2	0	7	4	0	2	3	4	3	2	3	2	3	1	0	0	2	0	1	1
child	6	3	1	1	0	1	0	0	0	2	0	1	2	0	2	0	3	4	1	4	0	0	2	7	1	1	1	1	0
long	2	1	2	1	1	1	2	0	0	3	1	1	0	1	1	4	2	1	4	1	0	7	2	1	1	0	0	0	1
man	2	4	2	1	3	1	3	0	1	0	1	0	1	1	5	6	3	4	1	2	7	3	3	4	0	1	0	0	1
woman	1	1	1	0	2	0	0	0	0	0	0	0	0	0	7	0	2	0	0	7	0	0	2	0	0	0	0	0	0
time	1	0	11	3	1	2	3	1	0	0	0	1	3	0	5	5	3	0	7	3	0	6	2	2	0	5	0	0	1
heart	10	0	0	1	1	1	1	0	0	2	6	2	0	0	4	1	1	2	2	3	2	2	2	1	0	2	0	4	0
poor	7	10	2	2	2	3	1	0	0	9	2	0	2	0	6	2	0	7	1	4	0	2	1	2	0	2	1	0	0
god	5	0	0	2	0	0	1	4	0	1	8	3	0	1	0	3	0	1	1	0	0	0	0	0	0	0	0	0	0
put	0	8	0	0	1	0	2	0	0	2	0	2	2	0	4	1	2	1	0	1	0	5	2	2	0	4	0	0	0
twist	1	3	1	0	0	2	0	0	0	0	0	0	8	0	4	0	0	1	0	0	0	0	0	1	0	0	0	0	0
rose	7	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	4	0	0	0	0	4	0
thief	0	0	0	0	0	0	0	0	0	0	0	0	1	7	0	2	1	0	1	1	0	1	0	0	0	0	0	0	6

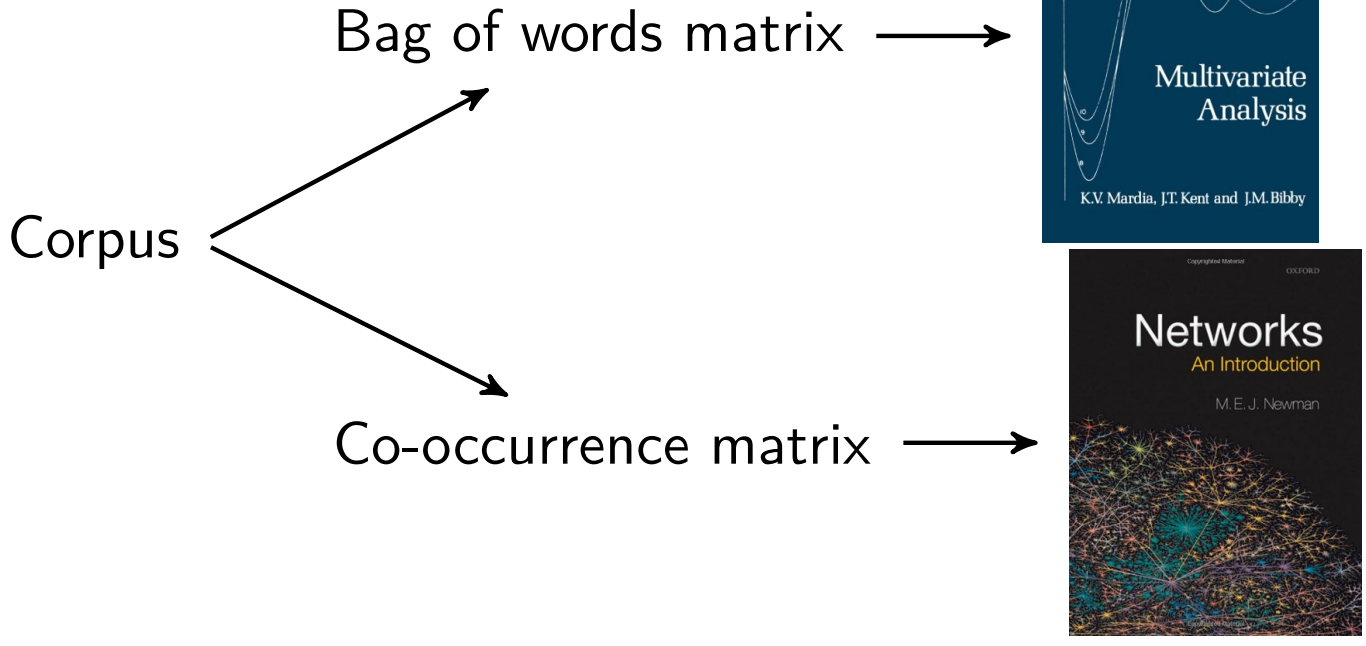
Speech from Oliver Twist: network visualisation

- Such matrix $X^{\text{word} \times \text{word}}$ can be identified with a "graph" (network).
- Lots of methods available for graphs.
- $\rightarrow$ Yves' talk later



Mathematical representation, X

Analysis, f(X)



Challenges and directions

- How to analyse time structured corpora? (E.g. newspaper archive)
  - Bag of words approach: each row of X is associated with a time t_i, then consider time-weighted $X(t) \rightarrow$ Anthony's talk.
- How to harness tools of network analysis to analyse co-occurrence networks, e.g. clustering? $\rightarrow$ Yves' talk.
- How to study time dependent networks? $\rightarrow$ ongoing work.

Summary

- All methods of corpus analysis are a function f(X) of a mathematical representation, X, of the corpus.
- Identifying X explicitly is helpful
  - to understand what information is used and what is discarded,
  - because abstraction provides a toolbox of methodologies, f(X),
- . . . and essential
  - to perform calculations for f(X) efficiently,
  - to develop new methodology, extending existing f(X).

Summary

- All methods of corpus analysis are a function f(X) of a mathematical representation, X, of the corpus.
- Identifying X explicitly is helpful
  - to understand what information is used and what is discarded,
  - because abstraction provides a toolbox of methodologies, f(X),
- . . . and essential
  - to perform calculations for f(X) efficiently,
  - to develop new methodology, extending existing f(X).
- Many promising directions ahead!