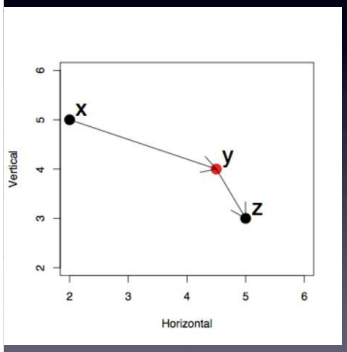
Lecture 2

Application to Textual Narrative

Fionn Murtagh

Triangular inequality holds for metrics



Example: Euclidean or "as the crow flies" distance

$$d(x,z) \leq d(x,y) + d(y,z)$$

Basic ideas and definitions

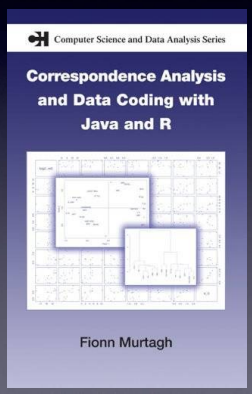
- Euclidean geometry for semantics of information.
- Hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. A further useful case is when the hierarchy respects chronological or other sequence information.

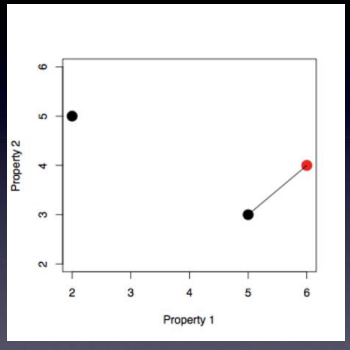
Ultrametric

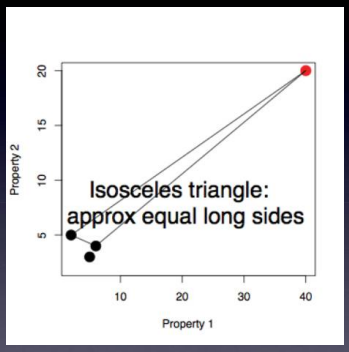
- Euclidean distance makes a lot of sense when the population is homogeneous
- Ultrametric distance makes a lot of sense when the observables are heterogeneous, discontinuous
- Latter is especially useful for determining: anomalous, atypical, innovative cases

Correspondence Analysis is A Tale of Three Metrics

- Chi squared metric - appropriate for profiles of frequencies of occurrence
- Euclidean metric, for visualization, and for static context
- Ultrametric, for hierarchic relations and for dynamic context



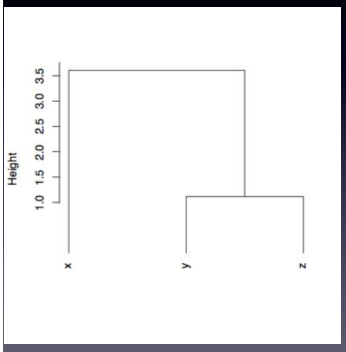




Analysis of semantics:
2. Hierarchy tracks anomaly and change

- Euclidean distance makes a lot of sense when the population is homogeneous
- Ultrametric distance makes a lot of sense when the observables are heterogeneous, discontinuous
- Latter is especially useful for determining: anomalous, atypical, innovative cases

Strong triangular inequality, or ultrametric inequality, holds for tree distances



$$d(x,z) \leq \max\{d(x,y), d(y,z)\}$$

$$d(x,z) = 3.5$$
$$d(x,y) = 3.5$$
$$d(y,z) = 1.0$$

Closest common ancestor distance is an ultrametric

Hierarchy, as well as geometry (Euclidean factor space as in Correspondence Analysis) for both understanding and working in complex systems.

In this course: applications to search and discovery, information retrieval, clusterwise regression, knowledge discovery.
Then analysis and synthesis of narrative, using filmscript and literary texts.
Also social media – Twitter.

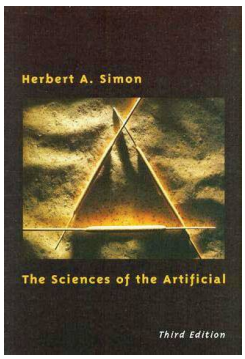
- Following slide is of: Herbert A Simon (1916-2001), Nobel Prize in economics 1978. He coined the terms: "bounded rationality", "satisficing", - and hierarchy as the architecture of complex systems. See: *The Sciences of the Artificial*, MIT Press.

Analysis of semantics:
1. Context - the collection of all interrelationships

- Euclidean distance makes a lot of sense when the population is homogeneous
- All interrelationships together provide context, relativities - and meaning

Chapter titles include:

- The psychology of thinking
- Remembering and learning
- The science of design
- Social planning
- The architecture of complexity: hierarchic systems



MIT Press, 3rd edn., 1996

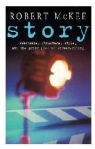
Analysis (and Support for Synthesis) of Narrative

- Casablanca movie - middle scene 43. Mapping and tracking emotion.
- Short look at: stochastic analyses of structure and style.
- Application that came from this work: supporting collective, collaborative narrative construction: book writing.
- After that: Social media – Twitter. New approach for assessing effectiveness.

Movie - Casablanca shot by Warner Brothers between May and August 1942

EXT. BLACK MARKET - DAYAt the linen stall, Ilsa examines a tablecloth which an Arab vendor is endeavoring to sell. He holds a sign which reads "700 francs."ARABYou will not find a treasure like this in all Morocco, Mademoiselle. Only seven hundred francs. Rick walks up behind Ilsa.

For McKee, composition of Casablanca is "virtually perfect".



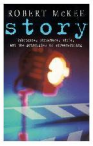
RICKYou're being cheated.She looks briefly at Rick, then turns away. Her manner ispolitely formal.ILSAIt doesn't matter, thank you.

- Scene 43 in Casablanca (out of 77 scenes).
- Crucial mid-point scene. Following McKee, I will analyze 11 subscenes ("beats").
- Right, first three subscenes (in blue, brown, red).

ARABAh, the lady is a friend of Rick's? For friends of Rick we have a small discount. Did I say seven hundred francs? You can have it for two hundred.Reaching under the counter, he takes out a sign reading "200 francs", and replaces the other sign with it.RICKI'm sorry I was in no condition to receive you when you called on me last night.ILSAIt doesn't matter.ARABAh, for special friends of Rick's we have a special discount. One hundred francs.He replaces the second sign with a third which reads "100francs."

Movie Casablanca by Warner Brothers between May and August 1942

For McKee, composition of Casablanca is "virtually perfect".

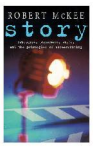


Analysis of Casablanca's "Mid-Act Climax", Scene 43 subdivided into 11 "beats" (subscenes)

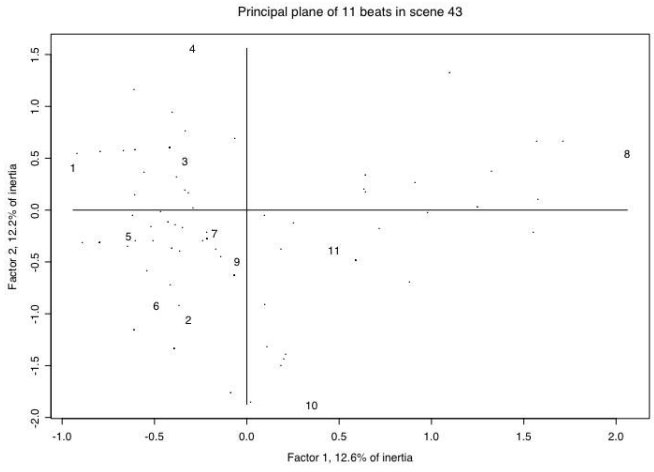
- McKee divides this scene, relating to Ilsa and Rick seeking black market exit visas, into 11 "beats"
- Beat 1 is Rick finding Ilsa in the market
- Beats 2, 3, 4 are rejections of him by Ilsa
- Beats 5, 6 express rapprochement by both
- Beat 7 is guilt-tripping by each in turn
- Beat 8 is a jump in content: Ilsa says she will leave Casablanca soon
- In beat 9, Rick calls her a coward, and Ilsa calls him a fool
- In beat 10, Rick propositions her
- In beat 11, the climax, all goes to rack and ruin: Ilsa says she was married to Laszlo all along. Rick is stunned

Movie Casablanca shot by Warner Brothers between May and August 1942

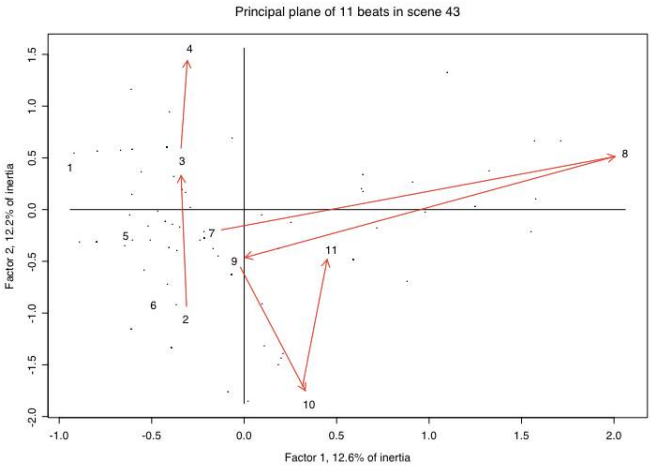
For McKee, composition of Casablanca is "virtually perfect".

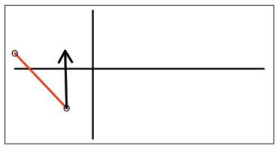


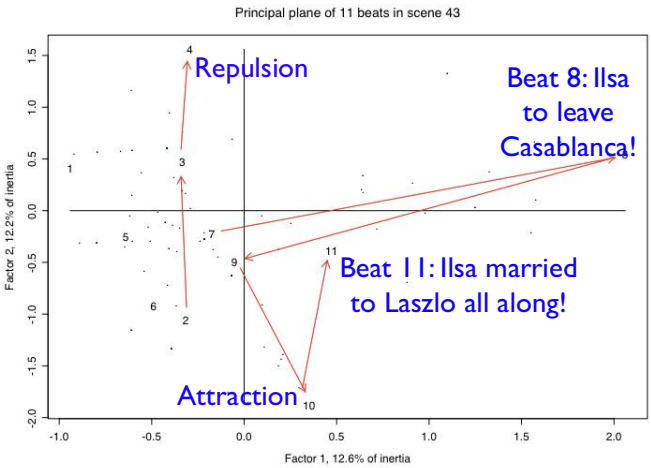
- Scene 43 in Casablanca (out of 77 scenes).
- Crucial mid-point scene. Following McKee, I will analyze 11 subscenes ("beats").
- Right, first three subscenes (in blue, brown, red).

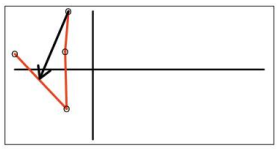


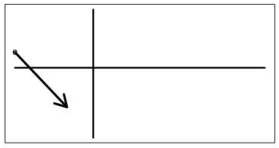
210 words used in these 11 "beats" or subscenes

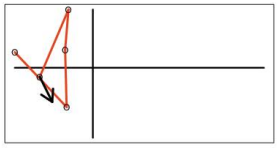

Principal plane of 11 beats in scene 43
Factor 2, 12.2% of inertia
Factor 1, 12.6% of inertia
1
2
3
4
5
6
7
8
9
10
11

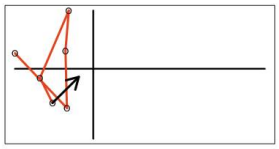


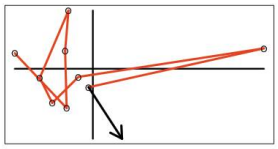

Principal plane of 11 beats in scene 43
Repulsion
Beat 8: Ilsa to leave Casablanca!
Beat 11: Ilsa married to Laszlo all along!
Attraction
Factor 2, 12.2% of inertia
Factor 1, 12.6% of inertia

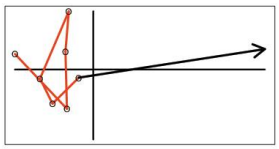


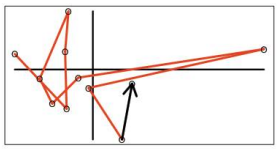


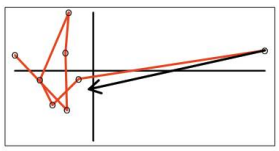












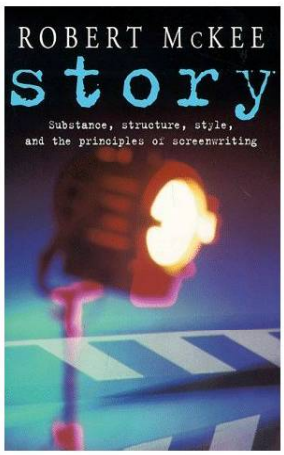
Foregoing example on YouTube (see www.narrativization.com)

- Back to a deeper look at Casablanca
- We have taken comprehensive but qualitative discussion by McKee (following slide) and sought qualitative and algorithmic implementation.
- For McKee: Text is the "sensory surface" of the underlying semantics.

McKee, Methuen, 1999

Casablance is based on a range of miniplots.

McKee: its composition is "virtually perfect"



Style analysis of scene 43 based on McKee Monte Carlo tested against 999 uniformly randomized sets of the beats

- In the great majority of cases (against 83% and more of the randomized alternatives) we find the style in scene 43 to be characterized by:
- small variability of movement from one beat to the next
- greater tempo of beats
- high mean rhythm

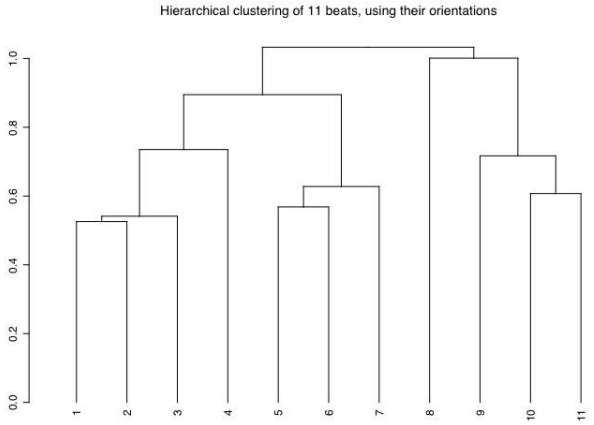
McKee's guidelines applied to Scene 43

- Lengths of beat get shorter leading up to climax: word counts of final five beats in scene 43 are: 50 - 44 - 38 - 30 --- 46
- The planar representation seen accounts for approx. 12.6 + 12.2 = 24.8% of the inertia, and hence the information
- We will look at the evolution of this scene using hierarchical clustering - but based on the relative orientations, or correlations with factors

Analysis of Narrative

Technical Issues Addressed

- We must consider complex web of relationships.
- Semantics include web of relationships - thematic structures and patterns. Structures and interrelationships evolve in time.
- Semantics include time evolution of structures and patterns, including both: threads and commonality; and change, the exceptional, the anomalous.
- Narrative suggests a causal or emotional relationship between events.
- A story is an expression of causality or connection
- Narrative connects facts or views or other units of information.



Full dimensionality analysis. Note caesura in moving from beat 7 to 8, and back to 9. Less so in moving from 4 to 5 but still quite pronounced.

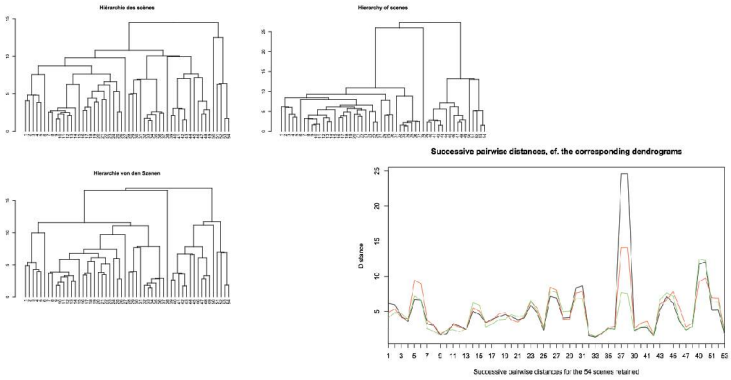
Our way of analyzing semantics with further applications

- We discern story semantics arising out of the orientation of narrative
- This is based on the web of interrelationships
- We examined caesuras and breakpoints in the flow of narrative
- With CSI (Crime Scene Investigation - Las Vegas - TV series) scripts: characterization

We find problems with word mapping between languages. Due to the mechanics of the translations, while we find emotion-relevant words in German ("dir", "deine", "dich", "du", and "sie", "sich") and in French ("je", "m" = "me", "aime", "tu", "t" = "te", and "je", "me"), we do not find much in English. Instead there is metadata in English such as "Speaking in French", "Speaking in German". Similarly for phrases, we have problems. E.g. "Here's looking at you, kid!" is mistranslated into "Ich schau dir in die Augen, Kleines!", and, in French, is "A vous, mon petit", and "A ta bonheur, mon petit".

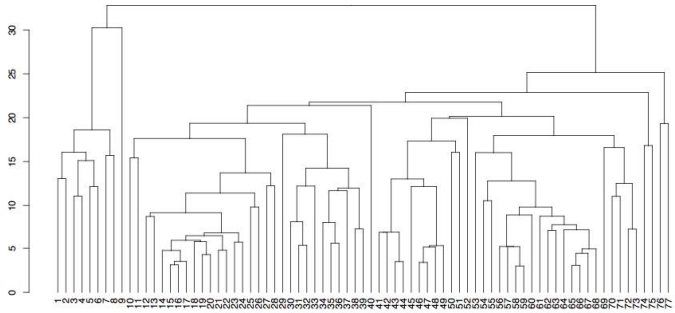
However semantic structural mapping works very well.

See hierarchies below. Bottom right: EN – black, DE – red, FR- green.

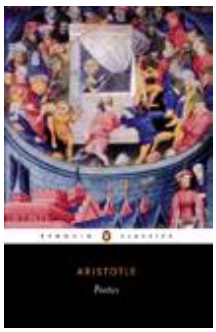


All Casablanca script: 77 scenes clustered - contiguity or sequence-constrained, complete link hierarchical clustering.

Shows up 9 to 10, and progressing from 39, to 40 and 41, as major changes.

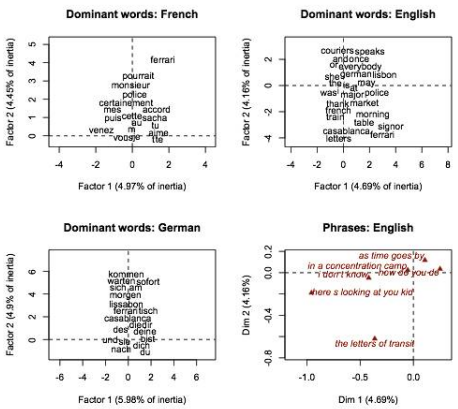


Text Synthesis



- Aristotle's Poetics (c. 350 BC)
- "Outlines and episodization" - "Stories ... should first be set out in universal terms ... on that basis, one should then turn the story into episodes and elaborate it."
- "... reasoning is the speech which the agents use to argue a case or put forwards an opinion"

Cross-language study: Casablanca subtitles in French, English, German. Respectively 1417, 1602 and 1476 subtitles. More than five occurrences of a word, no punctuation, upper case set to lower case: respectively retained 248, 331 and 226 words. But then scenes became empty: so a common set of 54 scenes was used.



Support environment for collaborative, distributed creating of narrative

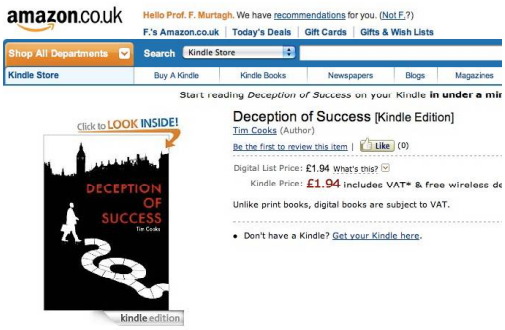
- Pinpointing anomalous sections
- Assessing homogeneity of style over successive iterations of the work
- Scenario experimentation and planning
- This includes condensing parts, or elaborating
- Similarity of structure relative to best practice in chosen genre

"Project TooManyCooks: Applying Software Design Principles to Fiction Writing"
Joe Reddington (RHUL, Comp. Sci.),

Doug Cowie (RHUL, English) and myself

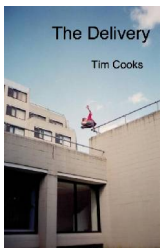
- Principal approach taken: Correspondence Analysis, principal plane essentially; and sequence constrained hierarchical clustering. In particular, the latter (using the full space dimensionality, and using the Euclidean metric, factor space, coordinates/ projections) is beneficial for a rapid view of internal, content-based structure.
- Joe Reddington also did some work for publishers in regard to newly submitted manuscripts.

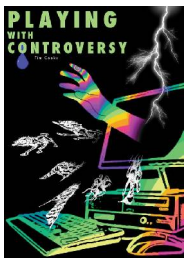
Collectively written in 4.5 days by a group of 10 children, of average age 12, in a school near London.

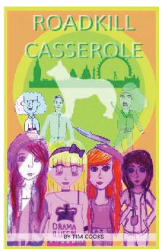


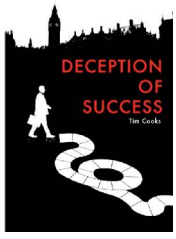
Social Media case study – Twitter

- Take some work of noted social / political science theorist, Jürgen Habermas
- Motivated by: Theory of Communicative Action (Theorie des kommunikativen Handelns)

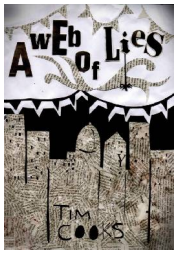








Books written collaboratively using support environment described here.
Upper left: RHUL English students; others: secondary school pupils.
Available for Kindle on Amazon.



1) Testing social media with the aim of designing interventions
2) Application here to environmental communication initiatives
3) Measuring impact of public engagement theory (in the Jürgen Habermas sense – public engagement centred on communicative theory; by implication therefore, discourse as a possible route to social learning and environmental citizenship)

1) Qualitative data analysis of twitter.
2) Nearly 1000 tweets between 1.10.2012 and 24.11.2012.
3) Evaluation of tweet interventions.
4) Eight separate twitter campaigns carried out.

Background and aims (those cited here were former colleagues)

- From Pianosi, Bull and Rieser: Quantitative social media (as Twitter here) measurement includes content indicators – most engaging content, top topics; and analysis of conversations created – length, number of people interacting, topic discussed.
- Pianosi, Bull and Rieser conclude: "… although useful in understanding the effectiveness of a communication campaign in its numerical terms, the proposed methodology can only be the first step of a more in-depth investigation about what people can learn during their on-line participation, and what is the perceived impact of the process on them, behaviour- or citizenship-wise. Consequently a more in-depth analysis of the characteristic of the community and a content analysis of on-line conversations is necessary …"
- **Our aims:**
- **Analyse the semantics of the discourse in a data-driven way. (Pianosi, Bull and Rieser: "top-down communication campaigns both predominate and are advised by those involved in "social marketing" …. However, this rarely manifests itself through measurable behaviour change …") Thus our approach is, in its point of departure and vantage point, bottom-up.**
- **Mediated by the latent semantic mapping of the discourse, we will develop semantic distance measures between deliberative actions and the aggregate social effect. We let the data speak in regard to influence, impact and reach.**

Transformed data used

Seq. no.	Tweet	Initiating – yes/ no	Campaign 1, 2, …, 8
1	Tweet 1	0	1
…	…	…	…
985	Tweet 985	0	8

The above shows the initial data set derived from the Twitter data spanning the eight campaigns. There are 985 tweets here.

Innovation in this work: how we address our objectives

- Impact: semantic distance between the initiating action, and the net aggregate outcome. This can be statistically tested. It can be visualized. It can be further visualized and evaluated.
- Essential enabling aspects are (1) the data structure input, comprising characterization of relevant actions, characterization of the initiating actions; and for all relevant actions, and the initiating actions, we have their context mode (called "campaign" here) which allows both intra and inter analyses. (2) Mapping of this characterization data (presence/absence, frequency of occurrence, mode category) into a semantic space that is both qualitatively (through visualization) and quantitatively analyzed. This sematic space is a Euclidean, factor space. For visualization we use 2-dimensional projection, but for quantitative analysis, we use the full factor space dimensionality, hence with no loss of information.

- Campaigns were as follows in the succession of tweets: 1 to 63; 64 to 133; 134 to 301; 302 to 409; 410 to 555; 556 to 730; 731 to 843; and 844 to 985.
- The initiating tweets for the eight campaigns are: 3, 65, 134, 303 and 304 (which were combined – the two taken together as one), 410, 557, 736 and 846.
- From all tweets, a set of, first, 3056 terms (see later slide for exact definition) were derived. Each tweet was cross-tabulated with those terms that were present for it. (Storage-wise, each tweet had 1 = presence, 0 = absence values for each of the 3056 terms. In some cases there were 2 or 3 presences.)
- The term set was reduced to 339 sufficiently often used terms. Some tweets thereby became empty, so the number of retained tweets became 968.

The 8 campaigns in late 2012

1.10-7.10: Climate change: The big picture and the global consequences

8.10-14.10: Climate change: The local consequences

15.10-22.10: Light and electricity

23.10-28.10: Heating systems

29.10-4.11: Sustainable Food choices

5.11-11.11: Sustainable Travel choices

12.11-18.11: Sustainable Water use

19.11-25.11: Sustainable Waste

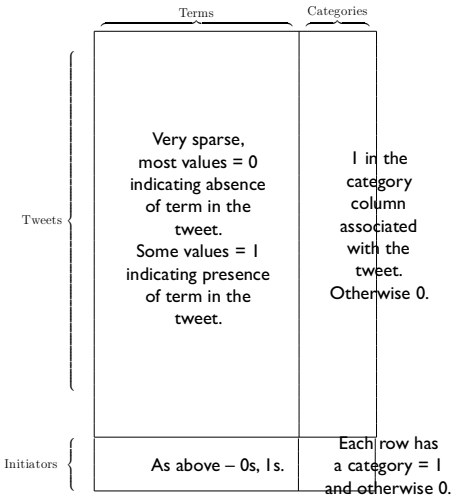
For the analysis, we distinguish between principal and supplementary rows (tweets) and columns (terms used by the tweets, and characterization in regard to the campaign).

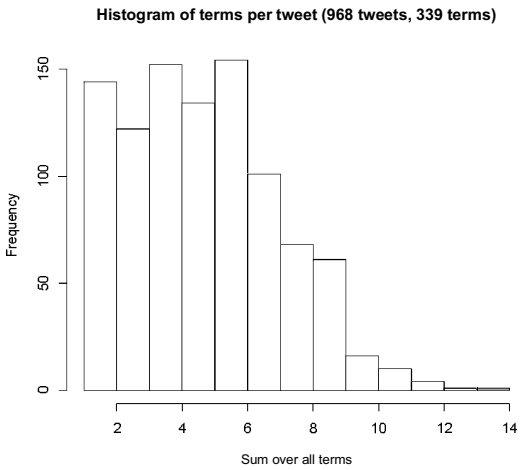
- The data to be analyzed then was:
- Principal rows: the set of 968 retained tweets, that do not include the initiating tweets.
- Supplementary rows: the set of 8 initiating tweets.
- Principal columns: the set of 339 terms retained.
- Supplementary columns: the set of 8 "indicators" for the 8 campaigns.

#	Term 1	…	Term 339	C1	C2	C3	C4	C5	C6	C7	C8
1	…	…	…	1	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…
968	…	…	…	0	0	0	0	0	0	0	1
3	…	…	…	1	0	0	0	0	0	0	0
65	…	…	…	0	1	0	0	0	0	0	0
134	…	…	…	0	0	1	0	0	0	0	0
303/4	…	…	…	0	0	0	1	0	0	0	0
409	…	…	…	0	0	0	0	1	0	0	0
556	…	…	…	0	0	0	0	0	1	0	0
735	…	…	…	0	0	0	0	0	0	1	0
845	…	…	…	0	0	0	0	0	0	0	1

Preprocessing the tweets x terms matrix

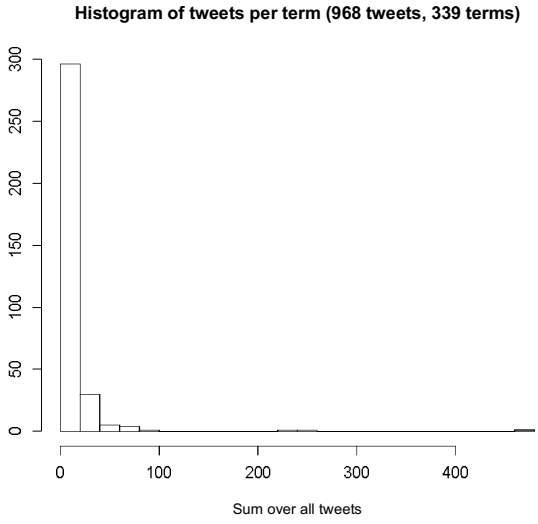
- The tweets x terms cross-tabulation is set up, with frequency of occurrence values. (The greatest frequency of occurrence value is 3. Typically the frequency of occurrence is 1. The cross-tabulation matrix is very sparse, with most values equal to 0.)
- Finally we require each set of presences of terms over all tweets to be at least 5, and also that the term is present in 5 tweets. (Very rarely used terms would not help our analysis.)
- The 968 retained (non-initiating) tweets, and the 8 initiating tweets, are crossed by 339 terms.
- Histograms of the tweet set, and of the term set, are shown in the following.
- The first histogram shows that up to 6 terms retained in a tweet is fairly constant in terms of frequency of such tweets. Tweets with more terms are fewer and fewer.
- The second histogram, of numbers of tweets per term, shows typical power law behaviour – many terms are present in a very small number of tweets; but as we look for terms in 100 tweets, in 200 tweets, in 300 tweets, we find very, very few cases of such tweets.





Preprocessing the set of terms, i.e. the content of the tweets

- Only alphabetic characters are retained. (So @, # are dropped but we can generally spot user or hashtag terms from the remaining term stump. Numerical data are dropped including dates, since we will focus exclusively on word-based data. Punctuation and special characters go too, e.g. in URLs. We could handle these were it advisable to do so.)
- The html expression for ampersand ("&"), in our processing left with a rump, "amp", is substituted with "and".
- All upper case is set to lower case (no loss of information involved).
- Deleted "ll" (from e.g. "I'll"), "s" (from e.g. "it's"), and "t" (from e.g. "isn't" or "wouldn't").
- We find 3323 terms used in the 985 tweets.
- Terms on a stopword list ("and", "the", etc.) are deleted (this decreases the 3323 term set to 3056 terms).



Our 8 campaign initiating tweets (for campaign 4 the two initiating tweets were merged together)

[C1] Introducing #climatechange! Is the climate changing?What are the observed changes?Are humans causing it? Discuss http://t.co/cMUOmbEt #dmuCC

[C2] Do you feel #climatechange is a distant issue? Read and listen to the climate witnesses in the UK http://t.co/FLWaTqTb

[C3] Goodmorning #DMU!! How was your weekend? Did you participate in the #marathon? We are talking about electricity this week! #dmuelectricity

[C4] Goodmorning #DMU!! How was your weekend? We are talking about gas and heating this week! #dmuenergy Wishing you all a nice #ecomonday!

[C4] Connect with us to discover what #DMU is already doing to cut its #gas use and tell us what you think we could all do to make it better!

[C5] Goodmorning #DMU!! We talk about #sustainable food this week. We have a question for you! What do you think does Sustainable Food mean?

[C6] Here I am, fueled with caffeine! This week we will be talking in particular of #transport. How do you get from home to #DMU? #dmutransport

[C7] New post! #Sustainable #Water | Are you familiar with the concept of #WaterSecurity? http://t.co/T9QYVITJ #DMU #climate #sustainabledmu

[C8] @SustainableDMU #MeatFreeMonday seems to have latched itself into my brain! Not a big meat eater but like having a dedicated veggie day!

- (Following slide)
- The planar display of factors 1 and 2.
- Tweets initiating campaigns, 1 to 8, are projected.
- The means or centres of gravity of the entire sets of (non-initiating) tweets for each of the 8 campaigns.
- We see that campaigns 3, 5, 8 have initiating tweets that are fairly close to the net overall campaign in these cases.
- The campaign initiating tweets, and the overall campaign means, are close to the origin, i.e. the global average. That just means that they, respectively – initiating tweets, and means – are relatively unexceptional, and express aggregates. (The factor 1, 2 scales inform us about that.) Note also the very low rates of inertia explained by the factors, again an aspect which is fairly standard for such analysis of very sparse cross-tabulations.

The terms retained for these particular initiating tweets, with frequency of occurrence. Note: for campaigns 1 through 8, we have summed frequencies of occurrence of terms: 4,4,7,14,10,6,7,5.

[C1] climate climatechange dmucc http (all 1)

[C2] climate climatechange http read (all 1)

[C3] dmu electricity goodmorning participate talking week weekend (all 1)

[C4] cut dmu dmuenergy ecomonday gas goodmorning heating nice talking tell week weekend (dmu, gas: 2; otherwise 1)

[C5] dmu food goodmorning mean question sustainable talk week (food, sustainable: 2; otherwise 1)

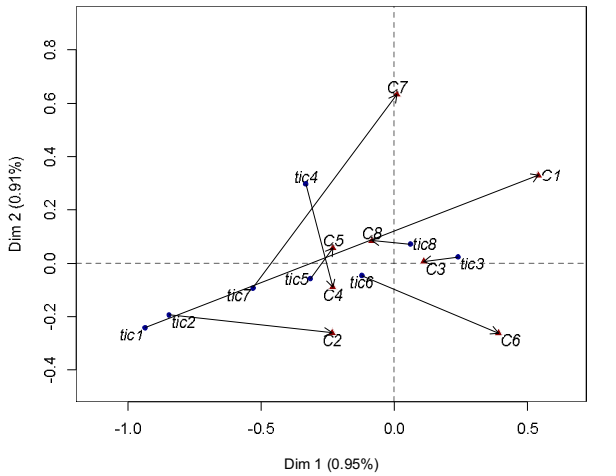
[C6] dmu dmutransport home talking transport week (all 1)

[C7] climate dmu http post sustainable sustainabledmu water (all 1)

[C8] day meat meatfreemonday sustainabledmu veggie (all 1)

Note that the campaign 4 tweet was a merged one (from original tweets 303, 304). In campaign 4, note that "gas" is both word and hashtag.
It is fairly easy to go back to the original tweets (previous slide) and see the hashtags, or the tweeters.
We keep "http", although a URL originally (see previous slide), but it tells us that more information – the web address – is in the tweet.
One conclusion we draw: low (retained) term initiating tweets (here in particular for campaigns 1, 2) are not plentiful in regard to information content.

8 campaign initiating tweets, and centres of gravity of 8 campaigns



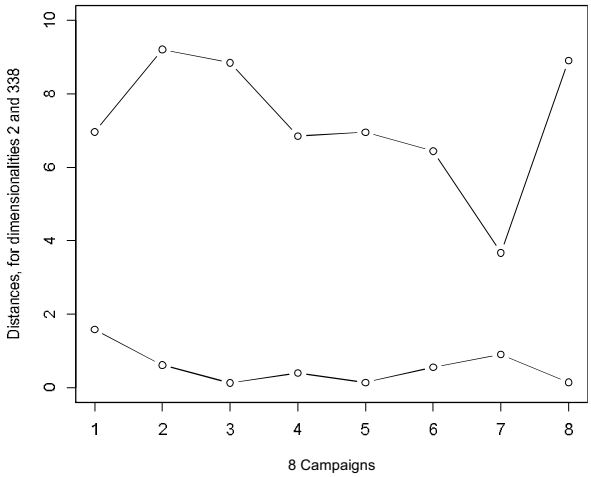
Correspondence Analysis

- Factors – in increasing order of importance, they provide latent semantic components.
- Analysis is carried out on the principal rows, columns. Then the supplementary rows, columns are projected into the analysis.
- Each term is at the centre of gravity of "its" tweets. Each tweet is at the centre of gravity of "its" terms.
- The factor space is a semantic space in that it takes account of all interrelationships – between all tweets, between all terms, between all tweets and all terms.
- Typically we visualize this semantic, factor representation of the data by taking two factors at a time. Hence, planar projections.
- In the analysis discussion to follow, we tidy up these displays, in order to highlight useful and/or important outcomes.

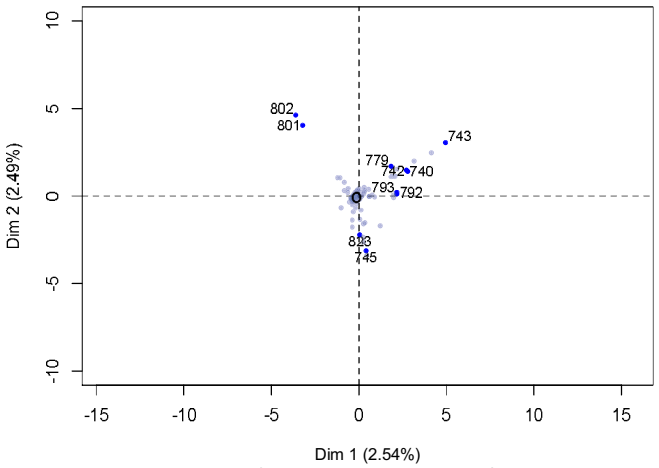
Impact of initiating tweet, quantified by its distance to its campaign mean (see following plot)

- While tweets initiating campaigns 3, 5, 8 are the closest to their respective campaign means, this is based on the best fitting planar, two-dimensional dimensions. It is based on the best factor plane, defined by factors 1 and 2.
- But the entire semantic space is of dimensionality 338. Looking at the distances between tweets initiating campaigns 1 to 8, relative to their respective campaign means (all in the factor space of dimensionality 338), we find a different (and more complete) perspective, where campaign 7 shows the most impact by its initiating tweet, followed by campaigns 6, then 4, then 5, then 1. Increasingly less impactful are 3, 8 and 2.

Distances between initiating tweet and campaign mean



Campaign 7: Factors 1, 2, with 10 most contributing tweets

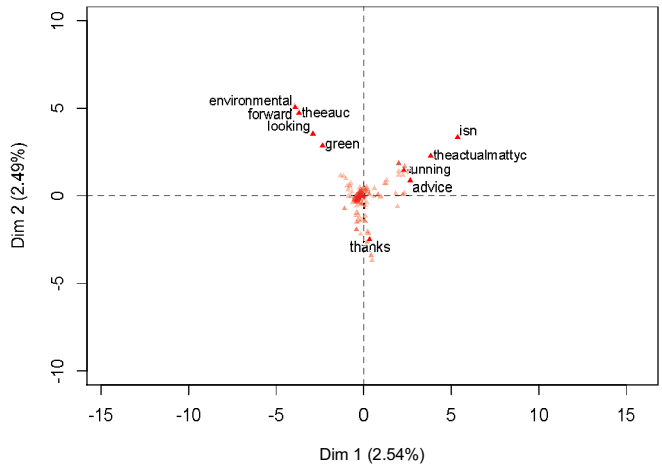


Initiating tweet: black "o". Otherwise sequence numbers of tweets in this campaign.

Statistical significance of impact

- The campaign 7 case, with the distance between the tweet initiating campaign 7, and the mean campaign 7 outcome, in the full, 338-dimensional factor (semantic) space equal to 3.670904.
- Compare that to all pairwise distances of non-initiating tweets. (They are quite normally distributed, with a small number of large distances.) The mean, mean – stdev, and mean – 2*stdev of these pairwise distances are: 12.64907, 8.508712, 4.368352.
- We find for campaign 7, the distance between initiating tweet and mean outcome, in terms of the mean and stdev of all (non-initiating) tweet, full dimensionality, pairwise distances, to be:
- mean – 2.168451*stdev
- For z = -2.16, the campaign 7 impact is significant at the 1.5% level (i.e. z - -2.16, in the two-sided case, has 98.5% of the Gaussian greater than it in value).
- In the case of campaigns 1, 4, 5, 6, we find them less than 90% of all pairwise distances.
- In the case of campaigns 3 and 8, we find them less than 80% of all pairwise distances.
- That only leaves campaign 2 as being the least good fit, relative to initiating tweet and outcome.

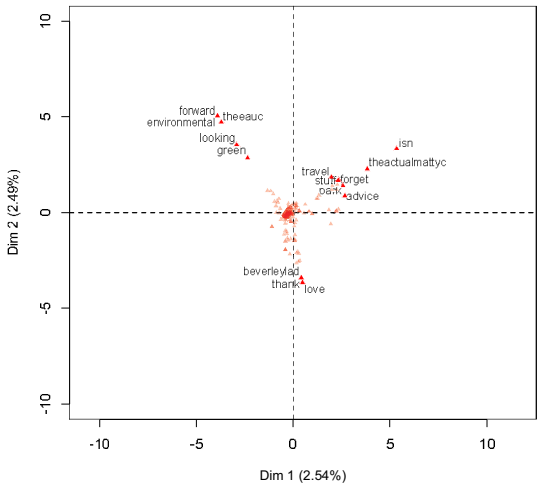
Campaign 7: Factors 1, 2, with 10 most contributing terms



Campaign 7

- Campaign 7: 12.11-18.11: Sustainable Water use
- Including the initiating tweet, there are 112 tweets (that have not become empty of terms in our term filtering preprocessing) in Campaign 7, and there are 176 terms that appear at least once in the set of tweets.
- First we show the Factors 1, 2 plane with the tweets, noting where the initiating tweet is located in this projection; and then we show the most important terms.
- Note @TheActualMattyC, @TheEAUC

Campaign 7: Factors 1, 2, with 15 highest coordinate terms



Conclusions 1/2: the Initiating Tweets relative to the Aggregate of Tweets in a Campaign

- We have traced out the semantic path from initiating tweet to the mean tweet of the associated campaign. We noted the differences between campaigns.
- We did this using the most salient – the most important – two-dimensional latent semantic, or factor, subspace, as well as in the full dimensionality space.
- We noted differences, e.g. Campaign 3 overall was closest to its initiating tweet in the two-factor projection; but with all information in use, Campaign 7 was the most effective campaign of all, in the sense of the initiating tweet being closest to the overall semantic mean of that campaign.

Conclusions 2/2: the Individual Campaigns

- In the eight campaigns we have pointed to what was more influential in terms of the underlying (latent semantic) components.
- In quite a few cases, this indicated who (the tweeter, @) and what themes (hashtag, #) were dominant. In other cases particular words were at issue.
- Our word set used was a carefully selected one. Nonetheless it was flexibly open (to various grammatical forms, and to stumps of words serving as proxies for words containing punctuation, web addresses, or other non-standard character strings).

Concluding comment on high dimensional analytics

- A fundamental aspect of the Twitter analysis was how a tweet, considered as a "campaign initiating tweet", differed from an aggregate set of tweets.
- The latter was the mean tweet, where the tweets were first mapped into a semantic space. The semantic space is provided by the factor space, which is endowed with a Euclidean metric.
- For very high dimensions, we find "data piling" or concentration. That is, the cloud of points becomes concentrated in a point. Now that could be of benefit to us, when we are seeking a mean (hence, aggregate) point in a very high dimensional space.
- A further aspect is when it is shown that the cloud piling or concentration is very much related to the marginal distributions.