Project Report

A decision support system for Info-graphics' designers:

Empirical study of Latent Dirichlet Allocation and Correlated Topic Models

Author: Meisam Hejazi Nia

Submitted to: Dr. Vibhav Gogate

Date: 12/08/2014

1. Abstract

The emergence of Web 2.0 revolutionized the content marketing strategies. Content marketing targets sales lead generation, and it takes various forms, such as text, image, and video. Info-graphic is a particular form of content marketing, and it is a type of representation and summary of the information to engage audiences with the content, and it uses audiences' eye processing capacity. As Edward Tufte suggests[1], human vision consists of millions of processing cells, and unlike the brain that can only keep 5 to 9 processes in short term memory at a time, human vision can process hundreds and even thousands of items simultaneously. This processing capability together with human need for insights make info-graphics an implausible tool for information elicitation from big data. Practitioners in industry have recognized this capacity of info-graphics, so they have made various attempts to give guidelines to potential marketers about how to design effective info graphics[2][3]. In particular, hubspot[4], an inventor of *inbound marketing* approach[5], has so far created more than 60 different info graphics to engage its potential consumers. Inbound marketing refers to marketing activities that bring visitors in, rather than marketers having to go out to get prospects' attention. Inbound marketing earns the attention of customers, makes the company easy to be found, and draws customers to the website by producing interesting content. In this study, we use a set of 260 info-graphics that we have collected from various websites including: Pinterest, hubspot, and informationisbeautiful.net, to quantify features that make an effective Info-Graphic. Our pilot project consists of 6 steps. To extract image information, in the first step we use RGB and HSV information of pixels of an info-graphic to create a vector of visual words. To extract the vector of visual words, we use an

[1] https://www.youtube.com/watch?v=g9Y4SxgfGCg
[2] http://blog.slideshare.net/2013/12/16/5-steps-to-creating-a-powerful-infographic/
[3] http://www.yeomansmarketing.co.uk/how-to-create-effective-infographics/2375/
[4] http://blog.hubspot.com/marketing/effectiveness-infographics
[5] http://en.wikipedia.org/wiki/Inbound_marketing

EM algorithm to identify five clusters in each image, and we build sorted histogram of the RGB and HSV information of each image. To extract text information, we also use an OCR combined with a dictionary process to extract text within the info-graphics. We first preprocess text data to remove stop words, and lemmatize the words. Then we use wordNet and Google's word2vec to find verbal similarity between the info-graphics. We merge both verbal and visual word vectors next and run two soft clustering methods, i.e. Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM) to cluster our info-graphics. We identified twelve different clusters of info-graphics. We named the clusters based on the word cloud of labels of info-graphics items within the clusters. Also based on non-parametric analysis we find that cool info-graphics about world's top issues and demographics has significantly higher social media hit than mobile and social media marketing info-graphics. In addition, info-graphics that contrast traditional and modern marketing approaches have significantly higher social media hits than other demographics. Interactive marketing info-graphics have significantly higher social media hits than social media marketing type info-graphics. From methodology standpoint, our approach gives a measure of the probability of membership in a cluster of viral info-graphics on each change that a designer makes. Our approach can be used as a decision support for info-graphic designer decisions, by benchmarking the new info-graphic design against cluster of viral info-graphics.

Keywords: info-graphics, soft clustering, topic model, Variational Bayes Methods, EM algorithm, VEM, LDA, CTM, k-mean.

1. Introduction

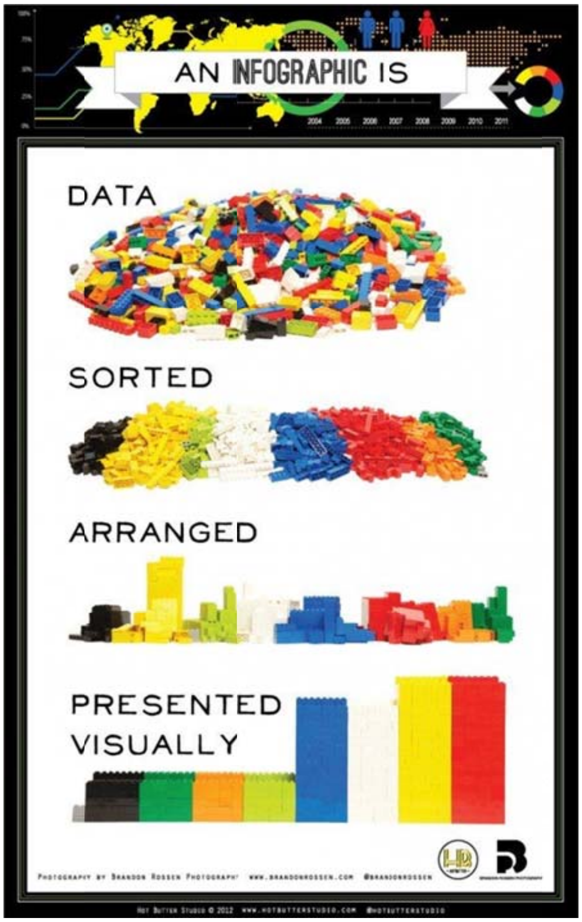
Image is a universal language. Images are both fun and amusing. Teaching theories suggest that different students have different ways of learning. Some students are more convenient in learning from text, while others are visual and learn more from pictures. Addressing the same problem, marketing practitioners have recently adopted the info-graphic approach to communicate more effectively. As Edward Tufte suggests[6], human vision consists of millions of processing cells, and unlike the brain that can only keep 5 to 9 processes in short term memory at a time, human vision can process hundreds and even thousands of items simultaneously. This human vision capability may suggest that images can be used more effectively if the marketing content not only includes text, but also it includes images. In addition, human eyes can extract insights and patterns faster if the data is more in pictorial forms. These characteristics together with the amusement capability of info-graphics have made info-graphics a popular tool for inbound social marketing.

Inbound social marketing is a word coined by hubspot, a British company. Inbound marketing is a way of promoting the company’s brand through blogs, podcasts, video, ebooks, enewsletter, white papers, search engine optimization, social media marketing, and other forms of content marketing to attract customers. We can call inbound marketing more as a modern type of marketing, in contrast to more traditional form that includes cold calling, direct paper mail, radio, TV advertising, sales flayer, telemarketing and traditional advertising, a type of marketing that is referred to with the name of “outbound marketing”. Perhaps the key distinction between inbound and outbound marketing can be find in the use of society as a medium for message. In other word, inbound marketers are interested to create quality content, because this quality content is

[6] https://www.youtube.com/watch?v=g9Y4SxgfGCg

more prone to be shared on the social media to become viral. This virility creates multiplicative effect for effort of inbound marketers. Inbound marketing earns user attraction rather than going out to get prospects' attention. Inbound marketers are more journalist and publisher type rather than traditional designers and marketers. These characteristics of inbound marketers make them more interested in pictorial methods, such as info graphics, rather than textual or simplified image methods.

Figure 1: An info-graphic of what is an info-graphic



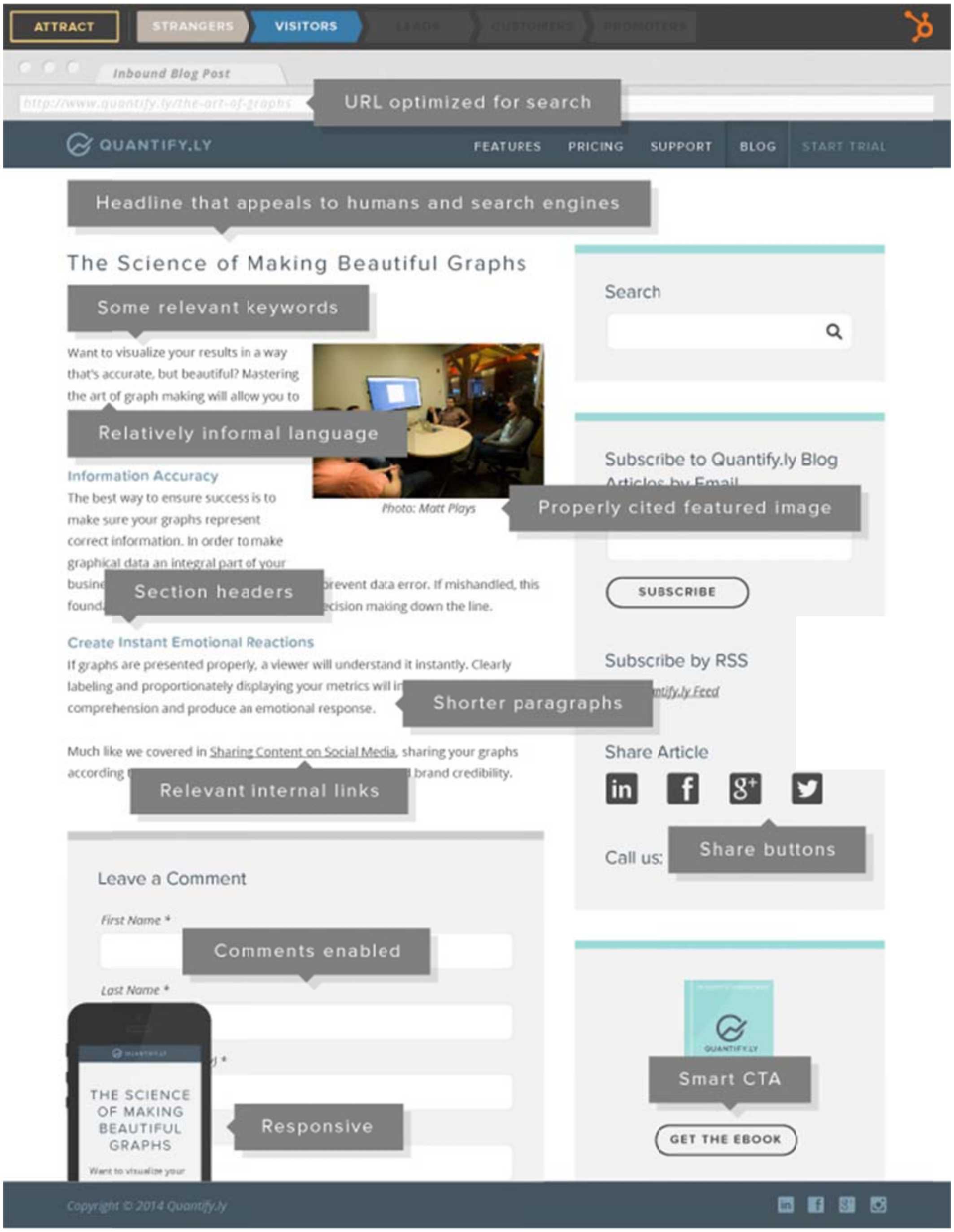
An info-graphic is a graphic visual representation such as a chart or diagram that is used to represent information, data, or knowledge intended to represent complex information quickly and clearly. Info-graphics can improve cognition by utilizing graphic to enhance human visual system's ability to recognize patterns and trends. The process of creating info-graphics can be referred to as data visualization, information design, or information architecture. Given these characteristics of info-graphics and the fact that they have not been studied quantitatively yet, we seek interesting patterns that can give us new information and guidelines on a type of info-graphic that is more prone to become viral. Figure 1, illustrates info-graphic summary of this paragraph, and figure 2 represents an info-graphic of a guide to create an effective blog post.

More particularly we asked, Does low level features of image give us meaningful and insightful classification of info-graphics? what is the optimal number of topics that we can categorize info-graphics to? Which features can give us more relevant result about topic classification of info-graphics? Do info graphics of different topics have systematically different level of virality from each other? Do models of correlated or uncorrelated topics of info-graphics fits our data better? Can we define a decision support system processes to evaluate whether each design choice of an info-graphic designers helps or hurts the possibility that an info-graphic becomes viral?

Answering these questions may help the info-graphic designers to develop more viral info-graphics. This viral info-graphics can help inbound marketers to meet their performance targets better. Our approach allows info-graphic designers to evaluate their info-graphics at each stage, and only accept the change that increases the probability that an info-graphic is from viral cluster of info-graphics. In other word, our approach can be used as a decision support for info-graphic designer decisions, by benchmarking the new info-graphic design against cluster of viral info-graphics. The underlying assumption to our approach is that although info-graphic as an art work is unique, yet there are some common features of info-graphics in a form of underlying patterns that makes an info-graphic pleasant and viral. This assumption may be backed up by practitioners' suggestions to piggy back on the successful art-works to help the new art-work become viral[7].

[7] http://blog.hubspot.com/blog/tabid/6307/bid/33611/7-Companies-That-Jumped-on-a-Viral-Craze-at-Just-the-Right-Time.aspx

Figure 2: an info-graphic guide to create an effective blog post



To meet these targets, we used a six step approach. In the first step to extract image information, we use RGB and HSV information of pixels of an info-graphic to create a vector of visual words. Our use of HSV information for clustering to the best of our knowledge is noble.

HSV stands for Hue, Saturation and Value color space, which is used more by designers compared with RGB space. To extract the vector of visual words, we use an Expectation Maximization (EM) algorithm to identify five clusters in each image, and we build sorted histogram of the RGB and HSV information of each image. To extract text information, we also use an OCR combined with a dictionary process to extract text within the info-graphics. We first preprocess text data to remove stop words, and lemmatize the words. Then we use wordNet and Google's word2vec to find verbal similarity between the info-graphics. The reason we used wordNet and word2vec was that the amount of text we identified in info-graphics is sparse. Therefore, to be able to use this sparse information, we aimed to map the vector of texts within each info-graphic into the broader domain of knowledge to extract similarity of sparse texts in each document to the others.

We then merge both verbal and visual word vectors and run two soft clustering methods, i.e. Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM) to cluster our info-graphics. The reason that we used soft-clustering LDA and CTM methods rather than hard-clustering methods such as k-mean or soft-clustering method such as Gaussian mixture method was grounded into the generalizability of the LDA and CTM methods to classify new data, based on the derived models, in addition to their theoretical ground. To be consistent, and show how these do methods cluster data different from the two focal methods of LDA and CTM, we presented the result of both Gaussian mixture and k-mean clustering.

We applied our approach to 355 info-graphic images we collected from informationisbeautiful.org, hubspot.com and pinterest.com. For each image we also collected number of social media activities, i.e. Facebook, Pinterest, Linkedin, Twitter likes shown on the website. Our approach can generally be classified as unsupervised machine learning approach to explore hidden patterns. Both CTM and LDA approaches create generative models from the data. The merit of these approaches together with Gaussian mixture soft clustering approach is their ability in summarizing big data (e.g. image of millions of pixels) to present useful patterns.

We find that image features can give us better clustering performance, compared with both image and text features of info-graphics. We identified twelve different clusters of info-graphics. We named the clusters based on the word cloud of labels of info-graphics items within the clusters. This approach may be credible because the basic guideline behind an info-graphic is that each info-graphic should have only one focal point. In addition, another basic guideline behind creating a viral info-graphic is to title the info-graphic, so that it is both relevant and representative. Therefore, a visual representation of titles of info-graphics within each cluster may give us a good idea of the content of each cluster.

To get better insight, we compared how different clusters virality is different across clusters, so based on non-parametric analysis we find that cool info-graphics about world's top issues and demographics has significantly higher social media hit than mobile and social media marketing info-graphics. In addition, info-graphics that contrast traditional and modern marketing approaches have significantly higher social media hits than other demographics. Interactive marketing info-graphics have significantly higher social media hits than social media marketing type info-graphics.

2. Problem Definition and Algorithm

2.1 Task Definition

We particularly ask the following questions: Does low level features of image give us meaningful and insightful classification of info-graphics? What is the optimal number of topics that we can categorize info-graphics to? Which features can give us more relevant result about topic classification of info-graphics? Do info graphics of different topics have systematically different level of virality from each other? Do models of correlated or uncorrelated topics of info-graphics fit our data better? Can we define a decision support system processes to evaluate whether each design choice of an info-graphic designers helps or hurts the possibility that an info-graphic becomes viral? Formally our problem can be defined as follows:

Input: Info-graphic images

Output: (1) A process to extract relevant features (2) A generative process that evaluates weather a new info-graphic is going to be viral

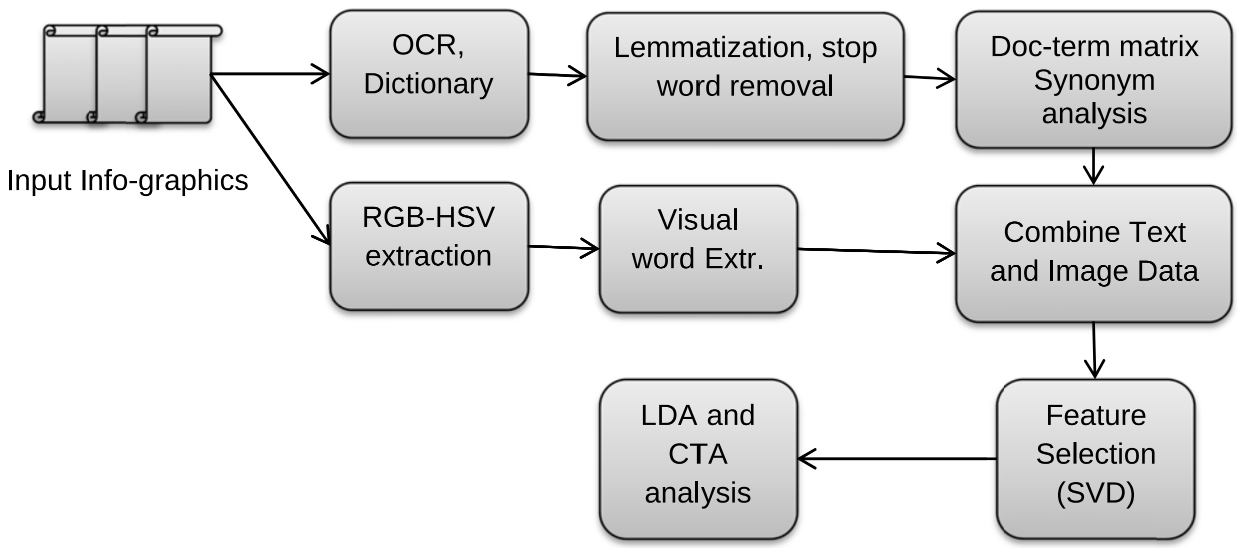
Practitioners list ten steps to create an effective info-graphic: (1) Gathering Data (2) Reading and highlighting facts (3) Finding the Narrative (4) Identifying problems (5) Creating a Hierarchy (6) Building a wireframe (7) Choosing a format (8) Determining a visual approach (9) Refinement and testing (10) Releasing it into the world[8]. Design is an iterative process, and it requires refinement and testing. Our approach uses machine learning approaches to extract collective wisdom of low level features (patterns) that create a viral info-graphic to guide designer in stage 8 and 9. In other word we attempt to quantify the art of user acceptance in info-graphic design, by extracting low level features of viral info-graphics.

2.2 Algorithm Definition

[8] http://www.fastcodesign.com/1670019/10-steps-to-designing-an-amazing-infographic

We start this section with defining our six step machine learning pipeline, or as we call it process of info-graphic-design decision support system. Figure 3 shows our machine learning pipeline.

Figure 3: Machine learning pipeline for info-graphic-design decision support system



We call this approach a six step approach as the visual word extraction process can be parallelized with lemmatization and stop word removal process, and RGB-HSV extraction process can be parallelized with RGB-HSV extraction process. In the first step we use Google's tesseract OCR which is an open source engine, yet as this engine extracts some irrelevant noises, we filtered its output with comparing each extracted keyword with a dictionary of English words. At the same time we extract RGB of each pixels of each info-graphic. We use a mapping between RGB and HSV to extract HSV information of each pixel. The size of this information about info-graphics are huge, so each info graphic's image is represented by millions of (R, G, B, H, S, V) hextuples. To extract visual words from this pictorial data, we adopted a k-mean algorithm, as suggested by Csurka et al. (2004) for classifying images. In summary, we first extract five clusters of (R, G, B, H, S, V) hextuples for each image. We pick not only mean of this hextuple for each image, and each cluster within each image, but also the size of the image in pixels and the density of each point within each cluster. We tried to use EM algorithm for this purpose, rather than k-mean, yet multimodal characteristic of the model and the size of our dataset, makes the Gaussian mixture model inappropriate for our purpose.

For text information as suggested by Grun and Hornik (2011), we removed the punctuation, numbers, stop words, and we lemmatized each keyword to its stem. Given these preprocessed text, we build the doc-term matrix for corpus of info-graphics, yet the result was really sparse. Therefore, to increase the quality of the data we used word thesaurus spaces such as wordNet and Google word2vec. We find similarity of all the keywords in one info graphic with all the keywords in another info graphic in these lexical databases of English words, and we used the average similarity across all keywords and across both databases to create a measure of similarity or distant of two info-graphic, and we used these similarity and distance measures as text features of each info-graphic. Our approach resembles the kernel approach proposed by Bishop (2006).

In the next step, we combined the full set of image and text features of info graphics into a single doc-term matrix. This matrix has 355 rows, i.e. for each document a separate row, and 392 columns. However, many of the elements of this matrix are zero, so the matrix is sparse. This scarcity suggested that we use dimensionality reduction techniques. As a result, we used Single Value Decomposition (SVD) method to keep 95% of the variation, and it gave us 30 new features. We coded the whole process until this point in python, and we used interfaces of WEKA for SVD, EM, and k-mean algorithms, and nltk interface for wordNet and word2Vec. Then we used R-interface of topic-models package to run CTM and LDA methods.

The reason we used LDA and CTM method rather than other soft clustering methods such as Gaussian mixture model or hard clustering methods such as k-mean or agglomerative clustering was that to design a decision support system, we needed a model not only based on theory, but also generalizable. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics (Blei et al 2003). Both LDA and CTM are generative approaches, and they use naïve conditional independence assumption, and they neglect the order of features by assuming exchangeability and using bag of words representation. These assumptions bring two main benefits to these approaches: simplicity, computational efficiency. Formally the LDA model assumes the following generative process for each item i in a collection C consisting of element (feature) e:

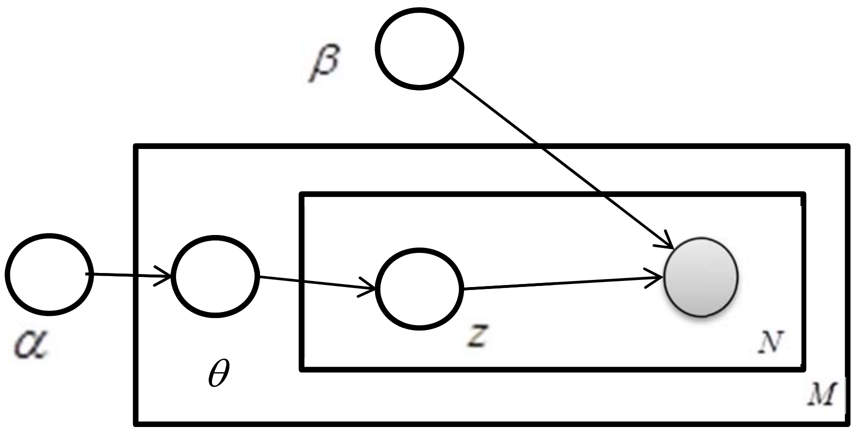
1. Choose N ~ Poisson (ξ), where N is the number of elements e
2. Choose $\theta \sim Dir(\alpha)$, where θ is the probability that a given document has primitive topic
3. For each of the N features i_n:
    a. Choose a topic $z_n \sim Multinomial(\theta)$
    b. Choose a feature i_n from $p(i_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic

A k-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex (a k-vector θ lies in the (k-1)-simplex if $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_{k1} - 1}$$

We represented the Probability Graphical Model (PGM) of LDA in figure 4. As figure depicts, there are three levels to the LDA representation. The parameters α, β are collection level parameters, and they are sampled once. The variable θ_d has Dirichlet distribution, and it is document level variable, so it is sampled once per document. This variable simply defines the weight distribution of topics within the document. Finally variables z_{d_n} and w_{d_n} are feature level parameters and they are sampled once for each feature within each document. Variable z_{d_n} defines the topic of n'ths word within document d, and variable w_{d_n} defines the feature instance that appears at location n within document d. As we can see an LDA model is a type of conditionally independent hierarchical model, and it is often referred to as parametric empirical Bayes model. One of the advantages of an LDA model is that it is parsimonious, so unlike probabilistic Latent Semantic Indexing (pLSI) model, it does not suffer from over fitting.

Figure 4: Graphical model representation of LDA



To estimate LDA model we define the likelihood of model in the following:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) (\prod_{n=1}^{N_d} \sum_{z_{d_n}} p(z_{d_n} \mid \theta_d) p(w_{d_n} \mid z_{d_n}, \beta) d\theta_d$$

The key inferential problem to solve for LDA is computing posterior distribution of topic hidden variables θ_d, z_d, the first one with Dirichlet distribution, and the second one with multinomial distribution. To normalize the distribution of words given α, β we marginalize over the hidden variables as following:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int (\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}) (\prod_{n=1}^{N_d} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} d\theta$$

Due to the coupling between θ and β in the summation over latent topics this likelihood function is intractable. Therefore to estimate it Blei et al. (2003) suggests using variational inference method. Variational inference or variational Bayesian refers to a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning. These family of methods are an alternative to sampling methods, and they are basically used to analytically approximate the posterior probability of the unobservable variables, in order to do statistical inference over these variables. These methods also give a lower bound to the marginal log likelihood. This family of lower bounds is indexed by a set of variational parameters. To obtain tightest lower bound we use an optimization procedure to select the variational parameters. A simple way to obtain a tractable family of lower bounds is to consider simple modifications of the original graphical model, by removing dependencies and introducing new variational parameters instead. In the LDA model we used following variational distribution to approximate posterior distribution of unobserved variables given the observed data s follows:

$$q(\theta, z \mid \gamma, \phi) = q_1(\theta \mid \gamma) \prod_{n=1}^{N} q_2(z_n \mid \phi_n)$$

Where $q_1(.)$ is a Dirichlet distribution with parameters γ and $q_2(.)$ is a multinomial distribution with parameters ϕ_n. Variational parameters are result of solving the following optimization problem:

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} D_{KL}(q(\theta, z \mid \gamma, \phi) \parallel p(\theta, z \mid w, \alpha, \beta))$$

where D_{KL} represents the Kullback-Leibler (KL) divergence between the variational distribution and the true joint posterior of latent parameters $p(\theta, z \mid w, \alpha, \beta)$. Formally, D_{KL} is defined as follows:

$$D_{KL}(q(\theta, z \mid \gamma, \phi) \parallel p(\theta, z \mid w, \alpha, \beta) = \sum_{(\gamma,\phi)} q(\theta, z \mid \gamma, \phi) \log\left(\frac{q(\theta, z \mid \gamma, \phi)}{p(\theta, z \mid w, \alpha, \beta)}\right)$$

As a result we can write KL-divergence in the following format:

$$Logp(w \mid \alpha, \beta) = L(\gamma, \phi; \alpha, \beta) + D_{KL}(q(\theta, z \mid \gamma, \phi) \parallel p(\theta, z \mid w, \alpha, \beta))$$

where

$$L(\gamma, \phi; \alpha, \beta) = E_q[\log p(\theta, z, w \mid \alpha, \beta)] - E_q[\log q(\theta, z)]$$

This relation suggests that maximizing the lower bound $L(\gamma, \phi; \alpha, \beta)$ with respect to γ and ϕ is equivalent to minimizing the KL divergence between the variational posterior probability and the true posterior probability. Expanding $L(\gamma, \phi; \alpha, \beta)$ using factorization of p and q gives the following:

$$\begin{aligned}
&L(\gamma, \phi; \alpha, \beta) = E_q[\log p(\theta \mid \alpha)] + E_q[\log p(z \mid \theta)] + E_q[\log p(w \mid z, \beta)] - E_q[\log q(\theta)] - E_q[\log q(z)] \\
&= \log \Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) \\
&- \log \Gamma(\sum_{j=1}^{k} \gamma_j) - \sum_{i=1}^{k} \log \Gamma(\gamma_i) + \sum_{i=1}^{k} (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}
\end{aligned}$$

Where $\Gamma(.)$ is gamma function and $\Psi(.)$ is its derivative. They key for this derivation is the following equation: $E[\log \theta_i \mid \alpha] = \Psi(\alpha_i) - \Psi(\sum_{j=1}^{k} \alpha_j)$, which is direct derivative of general fact that the derivative of log normalization factor with respect to the natural parameter of an exponential distribution is equal to the expectation of sufficient statistics. Collecting terms that are only related to each of the variational parameters γ and ϕ_{ni} from $L(\gamma, \phi; \alpha, \beta)$, and getting the derivative respectively give us an algorithm to solve the above optimization problem to find variational parameters. In particular, we can use simple iterative fixed-point method and update two variational parameters by the following equations until convergance:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log(\theta_i) \mid \gamma]\}$$
$$\gamma_i = \alpha_i + \sum_{n=1}^{n} \phi_{ni}$$

This optimization is document specific, so we view the Dirichlet parameter $\gamma^*(w)$ as providing a representation of a document in the topic simplex. In summary we have the following variational inference algorithm for LDA (Blei et al 2003):

(1) Initialize $\phi_{ni}{}^0 := 1/k$ for all i and n

(2) Initialize $\gamma_i := \alpha_i + N/k$ for all i and n

(3) **Repeat**

a. **For** n=1 **to** N

i. **For** i = 1 **to** k

1. $\phi_{ni}{}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i{}'))$

ii. Normalize $\phi_{ni}{}^{t+1}$ to sum to 1

b. $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$

(4) **until** convergence

This algorithm has the order of $O(N^2k)$. Given the variational Bayesian method we have tractable lower bound on the log likelihood, a bound which we can maximize with respect to α and β. We can thus find approximate empirical Bayes estimates for the LDA model via an alternating variational EM (VEM) procedure that maximizes a lower bound with respect to variational parameters γ and ϕ, and then, for fixed values of the variational parameters, maximizes the lower bound with respect to the model parameters α and β. The VEM algorithm is defined in the following:

1. (E-step) For each document, find the optimization value of the variational parameters $\{\gamma_d^*, \phi_d^* : d \in D\}$. This is done as described in the above variational inference algorithm.
2. (M-step) Maximize the resulting lower bound on the log likelihood with respect to the model parameters α and β. This corresponds to finding the maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step. The update for the conditional multinomial parameter β can be written out analytically as:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni}^* w_{dn}^j$$

The last concern about LDA is to make sure that sparsity does not make the likelihood zero, an extended graphical model with prior on β, where β is a k*V random matrix(k number of topics and V number of features, a row for each component), with independence identically Dirichlet distributed with parameter η rows assumption. Now β_i can be treated as a random variable to be endowed to the posterior distribution of hidden variables, giving us the following variational distribution with independence assumption:

$$q(\beta_{1:M}, z_{1:M}, \theta_{1:M} \mid \lambda, \gamma, \phi) = \prod_{i=1}^{k} Dir(\beta_i \mid \lambda_i) \prod_{d=1}^{M} q_d(\theta_d, z_d \mid \gamma_d, \phi_d)$$

To account for this modification, we only need to change the variational inference algorithm by augmenting the following update of variational parameter λ as follows:

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni}^{*} w_{dn}^{j}$$

This equation finalizes our plot of VEM algorithm to estimate an LDA model. There is an alternative approach proposed by Phan et al. (2008) that uses Gibbs sampling to estimate an LDA model. This approach draws from the posterior distribution of p(z|w) by sampling as follows:

$$p(z_i = K \mid w, z_{-i}) \propto \frac{n_{-i,K}^{(j)} + \delta}{n_{-i,K}^{(.)} + V\delta} \frac{n_{-i,K}^{(d_i)} + \alpha}{n_{-i,.}^{(d_i)} + k\alpha}$$

where z_{-i} is the vector of current topic memberships of all words without the i'th word w_i. The index j indicates that w_i is equal to the j'th term in the vocabulary. $n_{-i,K}^{(j)}$ gives how often the j'th term of the vocabulary is currently assigned to topic K without the i'th word, and the dot implies the summation over all relevant index instances. d_i indicates the document in the collection to which the word w_i belongs to. In this Bayesian formulation δ and α are the prior parameters for the term distribution of topics β and the topic distribution of documents θ, respectively. The predictive distribution of the parameter θ and β given w and z are given by:

$$\hat{\beta}_K^{(j)} = \frac{n_{-i,K}^{(j)} + \delta}{n_{-i,K}^{(.)} + V\delta} \qquad \hat{\theta}_K^{(d)} = \frac{n_{-i,K}^{(d_i)} + \alpha}{n_{-i,.}^{(d_i)} + k\alpha}$$

The likelihood for the Gibbs sampling also has the following form:

$$\log(p(w \mid z)) = k \log\left(\frac{\Gamma(V\delta)}{\Gamma(\delta)^V}\right) + \sum_{K=1}^{k} \{[\sum_{j=1}^{V} \log(\Gamma(n_K^{(j)} + \delta))] - \log(\Gamma(n_K^{(.)} + V\delta))\}$$

Finally, the correlated topic Model (CTM) has following differences from LDA method. First in the second step of the data generating process we have the following step substituted:

1. The proportions θ of the topic distribution for the document w are determined by drawing

$$\eta \sim N(\mu, \Sigma)$$

with $\eta \in R^{(k-1)}$ and $\Sigma \in R^{(k-1)\times(k-1)}$

Set $\tilde{\eta}^T = (\eta^T, 0)$. θ is given by: $\theta_K = \dfrac{\exp\{\tilde{\eta}_K\}}{\sum_{i=1}^{k} \exp\{\tilde{\eta}_i\}}$

This different definition results in the following log likelihood function:

$$p(D \mid \mu, \Sigma, \beta) = \log \int \{\sum_z (\prod_{n=1}^{N_d} p(z_{d_n} \mid \theta_d) p(w_{d_n} \mid z_{d_n}, \beta)\} p(\theta_d \mid \mu, \Sigma) d\theta_d$$

The variational distribution and optimization problem changes to the following:

$$q(\eta, z \mid \gamma, \phi) = \coprod_{K=1}^{k} q_1(\eta_K \mid \lambda_K, \upsilon_K^2) \prod_{n=1}^{N} q_2(z_n \mid \phi_n)$$

$$(\lambda^*, v^*, \phi^*) = \underset{(\lambda, v\ \ ,\phi)}{\arg\min}\, D_{KL}(q(\eta, z \mid \lambda, v^2, \phi) \parallel p(\eta, z \mid w, \mu, \Sigma, \beta))$$

As a result the VEM algorithm to estimate the CTM modifies as follows:

1. (E-step) For each document, find the optimization value of the variational parameters $\{\lambda_d^*, v_d^*, \phi_d^* : d \in D\}$. This is done as described in the above variational inference algorithm.

2. (M-step) Maximize the resulting lower bound on the log likelihood with respect to the model parameters μ, Σ and β. This corresponds to finding the maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step. The update for the conditional multinomial parameter β can be written out analytically as:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni}^{*} w_{dn}^{j}$$

3. Experimental Evaluation

3.1 Methodology

To evaluate the models we use log-likelihood to find the appropriate number of clusters. For the EM algorithm we use BIC criteria as it was the default of EM valuation software. We initialized LDA and CTM models multiple times and selected the maximum of likelihood across iterations. This because both LDA and CTM use VEM algorithm, which has EM algorithm in its core, and EM algorithm is prone to the problem of multiple modes. For k-mean method we used within sum of square to the total sum of square to find the optimal number of clusters. Our method was unsupervised learning, and we may be able to classify the clustering label of each info-graphic as dependent variable, and the text and image feature of our data as independent variables or features. We plan to collect new data and use it to validate our model in the next stage. As a result, so far our performance criteria are within sample, and it consists of likelihood, yet it is subject to over fitting. In the next stage also we will use portion of newly collected data as validation set for model selection.

3.2 Results

Table 1: Model selection based on (between sum of square / total sum of square)

Number of clusters (topics)	K-mean (image)	K-mean (full)
k = 3	66	96.9
k = 4	71.5	96.8
k = 5	73.5	96.7
k = 6	76.6	96.4
k = 7	78.7	95.1
k = 8	78.9	89.9
k = 9	80.8	87.8
k = 10	81.6	81.6

Table 2: Model selection based on log likelihood

Number of clusters (topics)	LDA (image)	CTM (image)	LDA-Gibbs (image)	LDA (full)	CTM (full)
k = 3	-275382	-275050	-198757	-1002505	-1002538
k = 4	-274685	-274268	-176261	-1002539	-1002589
k = 5	-274213	-273965	-161241	-1002561	-1002642
k = 6	-274680	-273779	-145156	-1002586	-1002631
k = 7	-274903	-273831	-133009	-1002610	-1002672
k = 8	-275274	-274193	-115634	-1002638	-1002712
k = 9	-275151	-273830	-99629	-1002660	-1002754
k = 10	-275382	-275050	-93164	-1002505	-1002538
k = 11			-88779	-1002539	-1002589
k = 12			-95304	-1002561	-1002642
k = 13			-97033	-1002586	-1002631
k = 14			-198757	-1002610	-1002672

Table 4: Model selection of Normal mixture model based on Bayesian Information Criteria (BIC) for model based clustering

Number of clusters	EII	VII	EEI	VEI	EVI	VVI	VEV
1	38286	38286	43511	43511	43511	43511	60367.86
2	41529	42845	43918	45172	45170	45857	61205.77
3	42498	43600	45053	46237	45630	46787	61409.09
4	42774	43971	46098	46492	46573	47132	61381.84
5	43110	44313	46400	47008	46701	47406	61207.54
6	43188	44572	46563	47257	46962	47571	60829.69
7	43259	44739	46659	47142	47083	47710	60545.68
8	43554	44803	46662	47412	47327	47591	60483.52
9	43518	44825	46758	47428	47360	47843	60162.49

"EII": spherical, equal volume
"VII": spherical, unequal volume
"EEI": diagonal, equal volume and shape
"VEI": diagonal, varying volume, equal shape
"EVI": diagonal, equal volume, varying shape
"VVI": diagonal, varying volume and shape
"VEV" = ellipsoidal, equal shape

Table 4: The clustering assignment based on maximum probability

	k-mean (image)	Gaussian Mixture (full)	LDA (image)	CTM (image)	LDA-Gibbs (image)	k-mean (full)	LDA (full)	LDA-Gibbs (full)
1	3	1	6	7	7	6	1	7
2	1	1	5	7	3	2	2	1
3	3	1	3	7	9	6	1	7
4	1	1	5	2	3	2	6	3
5	3	1	6	7	5	6	5	4
6	3	1	6	7	7	6	1	8
7	2	3	2	3	8	4	6	9
8	3	1	6	7	5	6	1	2
9	3	1	6	7	10	6	5	8
10	3	1	6	7	5	6	5	8
11	3	1	6	7	7	6	4	8
12	3	1	6	7	5	6	1	2
13	6	2	1	7	1	5	6	4

14	4	3	6	7	8	1	4	4
15	3	1	6	7	7	6	4	10
16	3	1	6	7	7	6	5	4
17	3	1	6	7	2	6	1	10
18	3	1	6	7	7	6	1	1
19	3	1	6	7	5	6	6	8
20	1	2	5	2	3	2	6	6
21	3	1	6	7	5	6	2	7
22	3	1	6	7	12	6	1	2
23	3	1	6	7	12	6	1	7
24	3	1	6	7	5	6	1	2
25	3	1	6	7	5	6	1	7
26	2	3	2	3	8	4	3	3
27	4	3	2	3	8	1	5	4
28	3	1	6	7	7	6	1	2
29	3	2	3	7	12	6	1	1
30	6	2	1	5	1	5	5	6
31	6	2	1	7	1	5	6	5
32	6	2	1	7	1	5	6	4
33	3	1	6	7	7	6	1	7
34	3	1	3	7	2	6	7	5
35	3	1	6	7	9	6	1	2
36	3	1	6	7	10	6	1	8
37	3	1	6	7	7	6	2	1
38	2	3	2	3	8	4	1	7
39	3	1	6	7	12	6	1	10
40	3	1	6	7	7	6	1	2
41	1	2	3	7	6	2	1	7
42	3	1	6	7	5	6	1	2
43	3	1	6	7	5	6	5	8
44	3	1	6	7	5	6	1	7
45	3	1	3	7	12	6	2	2
46	3	1	6	7	8	6	1	7
47	3	1	6	7	10	6	3	3
48	6	2	1	7	1	5	6	4
49	3	1	6	7	12	6	1	10
50	3	1	6	7	10	6	1	3
51	3	1	6	7	12	6	1	2
52	3	1	6	7	7	6	1	3
53	3	1	6	7	10	6	1	10
54	2	3	2	3	8	4	6	5
55	1	2	3	2	6	2	1	2
56	1	3	5	2	3	2	2	3

57	5	2	4	5	11	3	6	4
58	3	1	3	7	9	6	1	3
59	3	1	6	7	5	6	1	8
60	1	1	3	7	2	2	2	4
61	3	1	6	7	1	6	6	8
62	3	1	6	7	7	6	1	9
63	3	1	6	7	7	6	1	4
64	3	1	6	7	5	6	4	6
65	1	2	3	4	7	2	5	5
66	1	1	3	2	3	2	1	3
67	5	3	4	4	11	3	1	1
68	5	2	3	4	2	3	2	3
69	5	2	4	4	11	3	2	1
70	1	1	3	7	2	2	2	6
71	3	1	6	7	5	6	1	1
72	5	1	3	7	2	3	1	10
73	3	1	6	7	5	6	6	6
74	1	2	5	2	3	2	5	5
75	1	1	5	2	3	2	7	8
76	3	2	3	7	11	6	6	5
77	1	1	3	7	7	2	5	8
78	5	2	6	2	4	3	1	10
79	1	2	5	2	3	2	6	6
80	1	1	3	2	6	2	2	4
81	6	2	1	5	1	5	6	4
82	1	2	5	2	7	2	7	1
83	1	1	3	7	2	2	1	7
84	5	2	5	2	4	3	1	3
85	3	1	6	7	7	6	3	7
86	1	2	5	2	3	2	2	6
87	3	1	3	7	12	6	7	5
88	3	1	3	7	10	6	3	2
89	1	2	3	2	6	2	5	5
90	1	2	3	2	6	2	1	3
91	1	2	3	2	6	2	1	10
92	5	3	4	7	11	3	7	9
93	4	3	1	2	1	1	6	5
94	1	1	3	2	10	2	2	8
95	1	2	5	2	3	2	1	2
96	6	2	1	5	1	5	6	8
97	3	1	3	7	12	6	6	8
98	1	2	5	4	11	2	2	2
99	1	1	5	2	3	2	7	8

100	1	2	6	7	7	2	5	3
101	1	2	3	2	6	2	5	8
102	3	1	6	7	12	6	5	4
103	6	2	1	5	1	5	5	4
104	4	3	5	2	8	1	1	8
105	1	1	3	7	6	2	3	7
106	3	1	6	7	10	6	2	1
107	1	1	5	7	2	2	10	7
108	4	3	2	3	8	1	5	8
109	1	1	5	7	3	2	5	3
110	3	1	3	7	9	6	4	8
111	1	1	5	2	3	2	5	7
112	3	1	6	7	7	6	5	6
113	5	2	3	4	6	3	5	5
114	3	1	6	7	5	6	6	7
115	5	3	3	7	2	3	4	1
116	3	1	6	7	5	6	2	7
117	1	3	5	2	3	2	5	5
118	1	1	5	2	3	2	6	2
119	1	1	5	7	3	2	10	1
120	5	1	3	7	5	3	3	6
121	4	3	3	2	8	1	1	8
122	1	2	3	7	1	2	4	3
123	6	3	1	5	1	5	4	4
124	1	2	5	2	3	2	5	8
125	4	3	5	2	3	1	5	5
126	6	2	1	7	1	5	5	4
127	6	2	1	5	1	5	5	8
128	1	2	3	4	2	2	5	10
129	1	1	5	2	3	2	3	5
130	1	2	3	7	6	2	1	5
131	5	2	3	7	9	3	10	8
132	5	2	3	7	2	3	2	8
133	3	2	6	7	12	6	5	9
134	4	3	6	7	8	1	2	9
135	3	1	6	7	12	6	1	6
136	3	1	6	7	7	6	10	8
137	3	1	6	7	7	6	5	10
138	3	1	6	7	5	6	10	1
139	1	3	3	2	6	2	3	3
140	3	1	6	7	10	6	1	4
141	3	1	6	7	12	6	3	3
142	4	3	5	2	8	1	5	7

143	3	3	6	7	12	6	10	3
144	1	1	3	7	6	2	2	2
145	6	2	1	7	1	5	5	8
146	3	1	6	7	10	6	2	2
147	1	1	5	2	3	2	2	7
148	3	1	6	7	1	6	5	4
149	3	1	6	7	7	6	3	8
150	4	3	2	3	8	1	2	7
151	3	1	6	7	1	6	4	5
152	3	1	6	7	5	6	10	7
153	3	1	6	7	7	6	5	8
154	3	1	6	7	10	6	1	8
155	1	2	5	2	3	2	2	7
156	5	2	4	4	11	3	6	8
157	5	2	6	7	4	3	2	10
158	1	1	3	7	2	2	5	4
159	5	2	4	4	11	3	6	5
160	5	2	3	7	2	3	2	8
161	1	2	5	2	11	2	5	9
162	3	1	6	7	7	6	10	10
163	4	3	2	7	8	1	6	2
164	3	1	6	7	7	6	7	2
165	3	1	3	7	9	6	2	5
166	1	1	6	7	10	2	1	1
167	3	1	3	7	9	6	5	10
168	1	2	5	7	10	2	3	6
169	1	1	3	7	2	2	1	6
170	3	1	3	7	5	6	10	3
171	1	3	5	2	6	2	2	2
172	4	3	2	3	8	1	10	7
173	1	2	5	2	2	2	2	3
174	1	1	5	2	3	2	2	7
175	1	3	5	2	6	2	2	5
176	3	1	6	7	10	6	10	9
177	5	2	4	4	11	3	10	1
178	3	1	6	7	10	6	10	3
179	3	1	6	7	10	6	10	2
180	1	2	5	2	3	2	10	10
181	3	1	3	7	2	6	1	7
182	3	1	6	7	10	6	2	5
183	3	1	6	7	7	6	2	3
184	1	1	3	2	6	2	10	7
185	1	1	6	7	2	2	2	10

186	3	1	6	7	1	6	4	5
187	1	1	5	7	3	2	2	2
188	1	2	5	2	6	2	10	5
189	3	1	6	7	5	6	6	9
190	3	3	3	7	9	6	2	7
191	3	1	6	7	12	6	6	1
192	5	2	5	2	2	3	10	7
193	1	1	5	2	3	2	2	3
194	5	2	5	2	3	3	6	4
195	3	1	6	7	11	6	10	9
196	1	3	5	2	6	2	2	2
197	1	1	5	2	3	2	6	9
198	1	2	5	2	1	2	4	8
199	3	2	6	7	11	6	1	7
200	5	2	4	4	11	3	1	9
201	3	2	6	7	12	6	10	7
202	1	2	5	2	3	2	2	3
203	1	1	5	2	3	2	2	1
204	1	2	1	7	1	2	4	6
205	3	1	6	7	12	6	2	3
206	3	1	6	7	10	6	5	2
207	3	1	6	7	5	6	2	2
208	3	1	3	7	9	6	4	1
209	1	1	3	2	10	2	2	3
210	5	2	3	2	6	3	10	2
211	3	1	6	7	12	6	2	10
212	3	1	6	7	10	6	10	7
213	3	1	6	7	3	6	10	1
214	3	1	6	7	10	6	10	8
215	2	3	2	3	8	4	2	7
216	1	1	3	2	2	2	2	2
217	5	2	4	4	2	3	1	4
218	5	2	4	5	1	3	4	8
219	5	1	3	7	9	3	6	5
220	3	1	6	7	7	6	2	5
221	4	3	2	3	8	1	1	2
222	1	1	5	2	3	2	2	10
223	3	1	6	7	10	6	5	3
224	5	2	5	2	3	3	5	2
225	3	1	6	7	2	6	2	1
226	3	1	6	7	11	6	2	1
227	5	2	3	2	2	3	1	3
228	5	2	4	4	11	3	2	7

229	3	1	6	7	7	6	3	8
230	1	1	3	2	6	2	6	4
231	4	2	3	2	8	1	1	6
232	5	2	3	7	9	3	10	10
233	5	2	4	6	11	3	4	3
234	3	1	6	7	10	6	2	1
235	1	2	5	2	3	2	1	6
236	5	2	5	2	8	3	2	7
237	1	2	5	2	6	2	2	7
238	5	1	3	7	9	3	5	10
239	1	2	3	7	8	2	5	3
240	1	1	6	7	1	2	4	5
241	3	1	3	7	7	6	10	1
242	3	1	6	7	12	6	10	8
243	5	2	4	4	11	3	2	4
244	3	1	6	7	12	6	2	10
245	3	1	6	7	10	6	2	6
246	1	1	3	7	6	2	3	4
247	4	3	2	3	8	1	10	4
248	1	1	3	7	12	2	3	8
249	5	2	3	7	9	3	6	7
250	3	1	6	7	8	6	2	7
251	1	2	3	7	2	2	2	10
252	1	1	3	7	12	2	2	1
253	1	1	3	7	2	2	10	6
254	3	1	6	7	5	6	5	1
255	5	2	4	4	11	3	2	10
256	1	2	5	2	3	2	2	1
257	5	2	4	4	11	3	10	6
258	1	2	3	2	8	2	4	4
259	1	1	5	2	3	2	2	8
260	2	3	2	3	8	4	5	2
261	5	3	3	1	9	3	10	9
262	3	1	6	7	5	6	2	5
263	1	3	5	4	11	2	5	8
264	1	2	5	2	10	2	3	5
265	1	2	3	4	8	2	2	3
266	1	1	5	2	3	2	10	7
267	1	2	5	2	3	2	10	4
268	1	1	5	2	3	2	10	10
269	3	1	6	7	12	6	2	1
270	1	2	3	4	11	2	5	4
271	1	1	5	2	6	2	2	2

272	3	1	3	7	10	6	1	6
273	3	1	6	7	10	6	2	2
274	5	2	4	5	1	3	4	6
275	3	3	3	7	7	6	10	2
276	1	1	5	2	3	2	2	7
277	4	3	5	2	8	1	10	10
278	1	1	3	2	3	2	10	9
279	5	2	4	4	11	3	2	8
280	4	3	5	3	8	1	5	5
281	1	2	5	2	3	2	5	1
282	5	2	5	2	3	3	2	3
283	3	1	6	7	2	6	2	7
284	1	2	3	7	12	2	2	1
285	1	1	5	2	3	2	2	2
286	3	3	6	7	11	6	2	9
287	3	1	6	7	10	6	2	4
288	5	2	3	4	2	3	2	1
289	4	3	2	3	8	1	5	6
290	1	2	5	7	7	2	5	8
291	5	2	3	4	4	3	2	2
292	1	2	5	2	3	2	10	7
293	1	1	5	2	3	2	2	4
294	3	1	6	7	7	6	5	5
295	1	2	5	7	6	2	5	10
296	6	2	1	4	1	5	5	6
297	5	3	6	7	4	6	2	4
298	3	1	6	7	10	6	2	5
299	1	2	5	2	3	2	2	2
300	3	1	3	7	5	6	2	10
301	1	1	3	7	10	2	5	7
302	4	3	2	3	8	1	10	6
303	3	1	6	7	7	6	6	2
304	3	1	3	7	7	6	2	10
305	3	1	3	7	5	6	2	10
306	5	2	5	2	4	3	2	3
307	1	2	5	2	3	2	2	10
308	5	2	3	7	8	3	5	7
309	3	1	6	7	3	6	10	5
310	5	3	6	7	5	6	10	3
311	5	1	3	4	2	3	5	5
312	3	1	6	7	3	6	6	6
313	5	2	4	4	11	3	2	4
314	3	1	3	7	2	6	10	10

315	5	2	6	7	4	3	10	7
316	3	1	6	7	1	6	3	6
317	3	1	3	7	7	6	2	2
318	3	1	6	7	5	6	10	1
319	3	1	3	7	10	6	6	9
320	1	1	3	7	12	2	3	6
321	5	2	6	2	6	3	2	10
322	5	1	4	7	11	3	2	7
323	4	3	2	3	8	1	10	5
324	3	1	3	7	9	6	2	6
325	3	1	6	7	1	6	4	4
326	3	1	6	7	7	6	10	5
327	3	3	6	7	10	6	10	2
328	1	2	3	7	1	2	4	5
329	1	1	3	2	3	2	10	7
330	3	1	6	7	10	6	1	5
331	3	1	6	7	12	6	2	1
332	3	1	3	7	10	6	3	4
333	5	2	4	4	11	3	2	3
334	1	2	5	2	3	2	1	9
335	5	2	3	5	2	3	5	6
336	3	2	6	7	1	6	4	5
337	3	1	6	7	7	6	6	3
338	4	3	2	7	8	1	2	1
339	3	1	6	7	5	6	10	5
340	3	1	6	7	7	6	5	8
341	1	3	5	2	6	2	10	8
342	5	2	6	6	4	3	2	9
343	5	2	5	2	4	3	10	3
344	3	1	6	7	12	6	2	7
345	1	1	3	7	5	2	2	8
346	3	1	6	7	7	6	2	2
347	1	1	3	7	10	2	4	6
348	5	2	4	4	11	3	2	7
349	5	1	3	4	11	3	5	2
350	5	1	3	4	2	3	5	9
351	3	1	6	7	12	6	5	5
352	4	3	2	3	8	1	5	8
353	1	1	5	2	3	2	2	5
354	1	2	3	2	6	2	1	10
355	1	1	5	2	3	2	2	2

3.3 Discussion

Given our clustering results, we used word cloud to add meaning to the outcomes. To create a word cloud, for each cluster, we created a corpus of the titles of info-graphics. The basic idea behind this approach was that each info-graphic per definition is supposed to be created around a central main point. In addition, infographic creators select the title for their infographic meticulously to make sure that both it reflects its content, and it is general enough to be picked up as a relevant link by the search engines. Figure 5 illustrates word cloud of title of info-graphics within each cluster, and the clusters' names. After naming the cluster, we run a non-parametric t-test to compare whether social media activity (number of shares on Facebook, Pinterest, Linkedin, Twitter) as a measure of info-graphic virality differ systematically across the info-graphics clusters.

Based on non-parametric analysis we find that cool info-graphics about world's top issues and demographics has significantly higher social media hit than mobile and social media marketing info-graphics. In addition, info-graphics that contrast traditional and modern marketing approaches have significantly higher social media hits than other demographics. Interactive marketing info-graphics have significantly higher social media hits than social media marketing type info-graphics.

Our method has weaknesses that we have not made model selection based on hold-out validation set. In the next step we correct this defect.

Figure 5: Cluster titles’ word cloud for image analysis

<table>
<tr><td>Cluster 1: Cool info graphics about world’s demographic info-graphics</td><td>Cluster 2: Mobile and Buzz Design Info-graphics</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Cluster 3: Marketing design and Dashboard Info-graphics</td><td>Cluster 4: Face and Media Info-graphics</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Cluster 5: Traditional Marketing Info-graphics</td><td>Cluster 6: Social Media and Decision Making Info-graphics</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Cluster 7: General life Info-graphics</td><td>Cluster 8: Online professional design Info-graphics</td></tr>
</table>

Cluster 9: Responsive logos and brands Info-graphics

Cluster 10: International and online design Info-graphics

Cluster 11: Interactive Marketing Info-graphics

Cluster 12: Traditional vs. Online Media Info-graphics

Table 5: The basic statistics of clusters social media activity

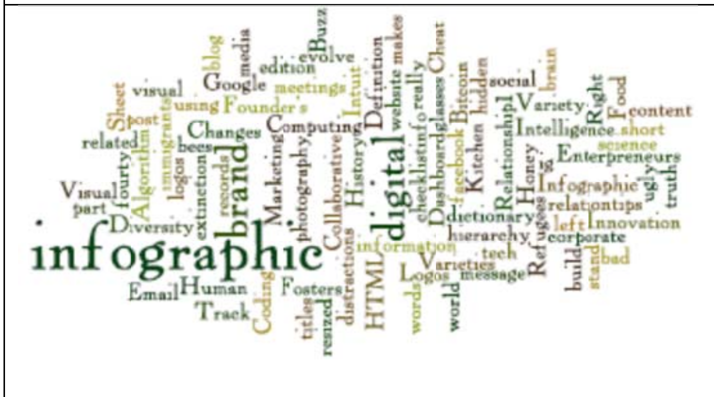
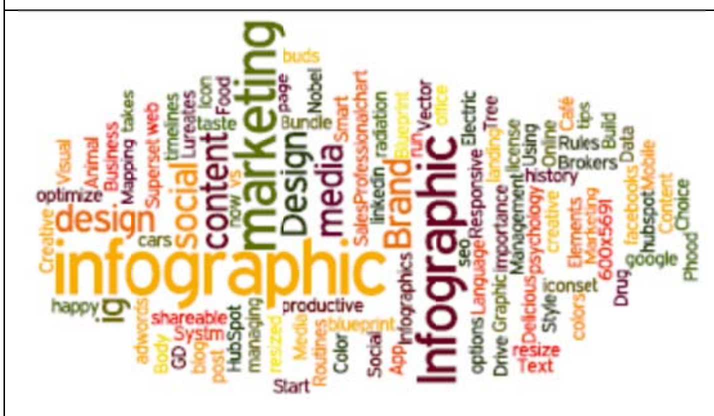
Cluster index	Cluster Name	frequency	average	variance
1	Cool info graphics about world's demographic info-graphics	28	2303.143	8744003
2	Mobile and Buzz Design Info-graphics	30	923.5333	1904941
3	Marketing design and Dashboard Info-graphics	53	1254.528	3987451
4	Face and Media Info-graphics	9	446.6667	350812.4
5	Traditional Marketing Info-graphics	31	2693.032	10011501
6	Social Media and Decision Making Info-graphics	26	960.1538	869841.9
7	General life Info-graphics	39	1774.615	5735747
8	Online professional design Info-graphics	33	1414.455	5010189
9	Responsive logos and brands Info-graphics	15	1194.6	3101275
10	International and online design Info-graphics	35	1354.057	6700740
11	Interactive Marketing Info-graphics	28	1030.571	5468611
12	Traditional vs. Online Media Info-graphics	28	1717.643	7377299

Table 6: Comparing Viral Measure (Social Media Activity) of clusters together by pairwise t-test (the first element represents the between group t-stat and the second element is corresponding t-test critical value)

	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10	cluster 11	cluster 12
cluster 1	(2.3,2)*	(1.89,1.99)*	(1.85,2.03)*	(-0.49,2)	(2.21,2)*	(0.81,2)	(1.33,2)	(1.33,2.02)	(1.36,2)	(1.79,2)	(0.77,2)
cluster 2		(-0.8,1.99)	(1,2.02)	(-2.81,2)*	(-0.11,2)	(-1.74,1.99)	(-1.04,2)	(-0.57,2.01)	(-0.82,2)	(-0.21,2)	(-1.42,2)
cluster 3			(1.2,2)	(-2.56,1.99)*	(0.71,1.99)	(-1.13,1.99)	(-0.34,1.99)	(0.11,2)	(-0.2,1.99)	(0.45,1.99)	(-0.87,1.99)
cluster 4				(-2.1,2.02)*	(-1.54,2.03)	(-1.64,2.01)	(-1.27,2.02)	(-1.22,2.06)	(-1.04,2.02)	(-0.73,2.03)	(-1.38,2.03)
cluster 5					(2.69,2)*	(1.38,1.99)	(1.88,2)	(1.7,2.01)	(1.89,2)	(2.27,2)	(1.26,2)
cluster 6						(-1.65,2)	(-0.97,2)	(-0.56,2.02)	(-0.74,2)	(-0.14,2)	(-1.35,2)
cluster 7							(0.66,1.99)	(0.85,2)	(0.73,1.99)	(1.27,2)	(0.09,2)
cluster 8								(0.34,2.01)	(0.1,2)	(0.65,2)	(-0.48,2)
cluster 9									(-0.22,2.01)	(0.24,2.02)	(-0.67,2.02)
cluster 10										(0.51,2)	(-0.54,2)
cluster 11											(-1.01,2)

* indicates whether the difference is significant with for 95% confidence interval

Figure 6: Cluster titles' word cloud for full analysis (for Gibbs LDA over the full set)

<table>
<tr><td>Cluster 1</td><td>Cluster 2</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Cluster 3</td><td>Cluster 4</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Cluster 5</td><td>Cluster 6</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Cluster 7</td><td>Cluster 8</td></tr>
</table>

	
Cluster 9	Cluster 10
	

4. Related Work

To the best of our finding this study is the first attempt to quantify the underlying features of viral info-graphics to build a decision support for info-graphic designers. From methodology standpoint, our approach gives a measure of the probability of membership in a cluster of viral info-graphics on each change that a designer makes. Our approach can be used as a decision support for info-graphic designer decisions, by benchmarking the new info-graphic design against cluster of viral info-graphics. This work relates to the work by Andrzejewki et al (2011), Bishop (2006), Blei et al (2003), Phan et al. (2008), and Hornik and Grun (2011) from methodological point of view. Although these works have suggested the use of LDA method and incorporation of kernels and domain knowledge at abstract level, our work tries to apply and integrate these approaches for the specific application of info-graphic design, and infographic designer decision support system. From domain point of view studies such as Ma et al (2004) have worked on infographics, yet the problem they try to address is different, and it is focused on info-graphic image downsizing. In addition, our work is related to work by Csurka et al. (2004) and Yang et al. (2007) in terms of use of bag of visual words for image classification, yet rather than running image classification, we used bag of visual words approach to run unsupervised clustering techniques that is generalizable to measure the potential of an info-graphic to become viral.

5. Future Work

Our study is only an initial attempt to unlock the black magic of art of creating the viral info-graphic. First from methodological stand point, we have to break the data into validation set and training set, and select the number of clusters based on the likelihood on validation set. Second, we can use hold out sample to compare our prediction about how viral a hold out info-graphic could be with the reality performance. We think we may be able to run bagging as a form of ensemble classifier to improve the potential bias of our method, as Dietterich (2000) suggests. Currently the image data performs much better than both image and text data together to extract meaningful clusters. In the next step we can hand craft the text feature of the infographics to extract more relevant features. We have not accounted for the age of the infographics in our non-parametric approach. In the next approach we can control this feature in our inference.

6. Conclusion

In this study, we use a set of 260 info-graphics that we have collected from various websites including: Pinterest, hubspot, and informationisbeautiful.net, to quantify features that make an effective Info-Graphic. Our pilot project consists of 6 steps. To extract image information, in the first step we use RGB and HSV information of pixels of an info-graphic to create a vector of visual words. To extract the vector of visual words, we use an EM algorithm to identify five clusters in each image, and we build sorted histogram of the RGB and HSV information of each image. To extract text information, we also use an OCR combined with a dictionary process to extract text within the info-graphics. We first preprocess text data to remove stop words, and lemmatize the words. Then we use wordNet and Google's word2vec to find verbal similarity between the info-graphics. We merge both verbal and visual word vectors next and run two soft clustering methods, i.e. Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM) to cluster our info-graphics. We identified twelve different clusters of info-graphics. We named the clusters based on the word cloud of labels of info-graphics items within the clusters. Also based on non-parametric analysis we find that cool info-graphics about world's top issues and demographics has significantly higher social media hit than mobile and social media marketing info-graphics. In addition, info-graphics that contrast traditional and modern marketing approaches have significantly higher social media hits than other demographics. Interactive marketing info-graphics have significantly higher social media hits than social media marketing type info-graphics. From methodology standpoint, our approach gives a measure of the probability of membership in a cluster of viral info-graphics on each change that a designer makes. Our approach can be used as a decision support for info-graphic designer decisions, by benchmarking the new info-graphic design against cluster of viral info-graphics.

References

Andrzejewski, D., Zhu, X., Craven, M., & Recht, B. (2011, July). A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (Vol. 22, No. 1, p. 1171).

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 1, p. 740). New York: springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1-15). Springer Berlin Heidelberg.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004, May). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (Vol. 1, No. 1-22, pp. 1-2).

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1-30.

Ma, R., & Singh, G. (2004, October). Large-scale infographic image downsizing. In *Image Processing, 2004. ICIP'04. 2004 International Conference on* (Vol. 3, pp. 1661-1664). IEEE.

Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web* (pp. 91-100). ACM.

Tirilly, P., Claveau, V., & Gros, P. (2008, July). Language modeling for bag-of-visual words image categorization. In *Proceedings of the 2008 international conference on Content-based image and video retrieval* (pp. 249-258). ACM.

Tufte, E. R. (2006). *Beautiful evidence* (Vol. 1). Cheshire, CT: Graphics Press.

Tufte, E. R. (1991). Envisioning information. *Optometry & Vision Science*,*68*(4), 322-324.

Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information* (Vol. 2). Cheshire, CT: Graphics press.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, *1*(1-2), 1-305.

Yang, J., Jiang, Y. G., Hauptmann, A. G., & Ngo, C. W. (2007, September). Evaluating bag-of-visual-words representations in scene classification. In*Proceedings of the international workshop on Workshop on multimedia information retrieval* (pp. 197-206). ACM.

Appendix A: the title of the info-graphics

Index	Title
1	620 Who rule the social web
2	1276 Diversity in tech
3	1276 Best in show
4	1276 Cash Crops 6thNov
5	1276 islamic sects Nov18 onroll
6	1276 Antibiotic Abacus july14
7	far future timeline
8	1276 influ venn za6
9	1276 Codebases
10	1276 microbescope4
11	1276 Common Mythconceptions Oct22nd
12	The Middle East
13	iib death wellcome collection fullsize
14	1276 Rape3
15	1276 Being Defensive
16	1276 punytive damages
17	1276 scale of devastation
18	1276 relationtips 3
19	1276 gigatons CO2 Oct14
20	1276 Rhetological Fallacies EN
21	1276 chicks rule
22	1276 who really runs the world
23	US film industry 1
24	1276 Taxonomy of Ideas1
25	1276 books everyone should read

26 selling out 550
27 1276 snake oil supplements Apr14
28 1276 left right usa
29 1276 hierarchy of digital distractions
30 1276 occupy wall st
31 1276 taste buds
32 HPV 2
33 1276 horoscoped
34 1276 radiation chart 2013
35 data info knowledge wisdom
36 breakups facebook
37 1276 Varieties of Human Relationship1
38 goggle boxes
39 in deeper water
40 1276 mountains molehills aug2014 22
41 1276 when sea levels attack Feb14
42 planes volcanos
43 1276 Articles of War
44 940 china censorship
45 1276 billion dollar o gram 2009
46 1276 drugs legalised
47 H1N1 550
48 1276 20121
49 good infodesign 550
50 1276 international number ones
51 1276 colours in culture
52 drug deaths 1 460
53 kyoto 550
54 1276 timelines
55 1276 billion pound o gram
56 1276 billion dollar o gram
57 1276 climate skeptics
58 1276 left right world
59 twitter2 550
60 1276 buzz v bulge
61 1276 reduce your chances
62 1276 Drugs world
63 nukes 550
64 1276 Common Mythconceptions Oct22nd
65 shareable social media infographic
66 work bffs
67 blog post titles
68 bounce rate (infographic)

69 visual content 1
70 SEO How to write content that ranks 2014 [infographic]
71 do this not that facebook edition (infographic)
72 HowtoOptimizeblog (infographic)
73 eye tracking
74 Data Brokers HubSpot Infographic
75 email marketing myths [infographic]
76 Mapping out facebooks options for blog [infographic]
77 Words That Convert Uberflip Infographic
78 The Nuts and Bolts of a Perfect Facebook Post 1
79 whatmakesagoodheadline
80 website design features IG
81 BTE infographic 4
82 ugly truth meetings ig
83 inbound blog
84 mixing typefaces infographic
85 famous rebrands
86 The Hidden Cost of a Failed Sales Manager
87 what is google adwords ig
88 How To Get More Blog Subscribers Infographic small
89 optimize landing page ig
90 inbound social
91 Purchase Decisions Infographic
92 what is responsive website design ig
93 33 linkedin tips infographic
94 psychology of color ig
95 great divide in content marketing ig
96 The Power of Visual Content infographic
97 value of coupons in digital marketing infographic
98 Holiday trends infographic
99 how your brain sees logos infographic
100 death of the office
101 12 Twitter Stats to Help Get You More Conversions (1)
102 What Makes Someone Leave Website
103 Qvidian Sales Playbooks Infographic
104 great american pumpkin takeover ig
105 typography and fonts infographic
106 short world records infographic
107 expiration date entrepreneur
108 Calls are the new Clicks Infographic
109 does email work create resentment infographic
110 brand logos with hidden messages
111 reduceoptionsincreaseconversions

112 2014 holiday shopping guide
113 seo then vs now
114 2014holidayshoppingguide600
115 141008 Intuit Bitcoin
116 social thankyou infographic 02
117 the history of marketing hubspot resized 600
118 infographic infographic resized 600
119 Visual History of Google Algorithm Changes
120 the power of visual communication infographic
121 7 Superpowers of a Knockout Infographic Socially Sorted
122 Post Pin Tweet Best Time Outreach
123 ranking factors infographic 2
124 personal branding infographic
125 social media design blueprint
126 Facebook Ad Infographic
127 Logo infographic
128 impactbnd inbound marketing process final resized 600
129 GD SalesProfessional Infographic resize (1)
130 managing content marketing infographic 600x5691
131 12 homepage elements hubspot infographic
132 State of sales productivity 2014 infographic final 1
133 blogging secrets
134 color purchases infographics
135 calculating customer LTV
136 so what is inbound marketing1 resized 600
137 Social Media Facts and statistics you need to know
138 checklistinfo resized 600
139 The blogconomy infographic 640x5604
140 Gigya Sharing Infographic Q3 2013 1
141 essential blog post ingredients infographic
142 5min LinkedIn Infographic Bluewire Media
143 foursquare2010 resized 600
144 pushing the e envelope (crop)
145 10 rules that make infographics effective cool and viral
146 Social Media in Business
147 Sample Infographics
148 Dress Daper
149 infographic element design
150 common octopos
151 Should I post this
152 creative idea
153 how steve job started
154 what is an infographic

155 smartphones
156 Graphic Design
157 Television history
158 iPhone evolution
159 Graphic Design
160 Research
161 What's your snack and food
162 Too late to learn success
163 The water rich vs. the water poor
164 What does your logo say about your business
165 Tree iconset
166 What makes an infographic bad
167 World Café
168 Vinyl record in 50 years
169 Typography and fonts
170 Galaxy comets
171 Visual content takes off
172 Global carbon foot print
173 Brand identity systm
174 29 ways to stay creative
175 Visual content takes off
176 Most educated Nations of the world
177 How are we using social media
178 Web design Trends
179 Cost of owning a pet
180 Adobe Keywboard
181 Interior design by decade
182 Delicious Café
183 The coffee facts
184 Firt man to attempt to walk the length of the nile
185 Travel infographic
186 The psychology of colors in marketing
187 My creative process
188 Animal Icon
189 Invitation to Wedding
190 Infographic creation process
191 Food related information
192 How we use our mobile devices
193 Exciting powerpoint presentation
194 Guide to starbox Espresso
195 ice-cream
196 Visual content takes off
197 ten powerful body language tips

198 Sweet and tooth
199 American Students Studying Abroad
200 Effective Story telling
201 The hobbit by numbers
202 Nature offers health benefit to our symptoms
203 Right and left part of brain
204 Future of Mobile Video
205 Ebola Disease
206 Why US hospitals are not as safe as you think
207 How to build the brand of Awesomeoness
208 fourty brand logos with hidden message
209 Hawaii
210 Earth information
211 Fourty ways to stay creative
212 Free rider
213 Track your glasses
214 International Oceanography
215 Geographical software application
216 Zambia Demographics
217 Paris London Urban Head to Head
218 online game
219 Responsive web design
220 Content Management Systm
221 Quantom
222 Sample Infographics
223 Evolution of an Entrepreneur
224 iOS and android design and guidelines
225 Honey bees extinction
226 The science of photography
227 Web designer developer
228 responsive design process
229 Londoner's guide to London
230 How to influence and persuade
231 The startup ecosystem predator vs prey
232 Critical Elements every home page must have
233 User interface future pack
234 What is digital Marketing
235 How to start a web design project
236 Online Brands
237 Don't Suck at meetings
238 A breif introduction to Typography
239 The Anatomy of a perfect website
240 How to be more creative

241 Should you build a website
242 The most popular free lancer tools
243 Dashboard of Collaborative Computing
244 The world happeiest and unhappiest countries
245 Java script
246 Social Media vs. Tradditional Media
247 An anatomy of a web designer
248 Social Media vs. Tradditional Media
249 Revolve Ecommerce Fashion site
250 How much did famous logos cost to design
251 Home page design of fortune 500 companies websites
252 How corporate Logos evolve
253 Company profile
254 Email Coding HTML
255 The world's weirdest festivals
256 Dashboard for Collaborative Computing
257 Infographic Vector Elements
258 Advertising vs reality
259 26 things to note before you develop a website
260 The global water crisis
261 The earth's oldest trees
262 Electric cars
263 Website design
264 How to run a productive and happy office
265 Logo design process
266 Dashboard for collaborative computing
267 Most important CSS3 properties
268 Infographic survey
269 Founder's dictionary, Enterpreneurs' Buzz words
270 Is your job killing you
271 Sample Infographics
272 RGB vs CMYK colors
273 HTML5: Past, Present & Future
274 A perspective on time
275 Nike wearable device
276 Good and Bad Habits
277 Basketball team League
278 15 Blurred Backgrounds
279 Modular interactive Element
280 Brand license to Drive
281 What does your brand stand for
282 The Internet, Topology of Autonomous Systems
283 Mobile Dashboard

284 What really Fosters Innovation
285 True Colors
286 Should I send this email
287 Planning design and optimising website simplified
288 Refugees and immigrants
289 How Bikes can Save Us
290 HTML 5 cheat sheet
291 Graphical Designer and Filmmaker profile
292 Autonomy of Pinterest Influencers
293 Sample Infographics
294 Build Your Brand Online
295 HTML 5 cheat sheet
296 World Demographics
297 Flat 3D Mockup Kit
298 Food and Drug
299 World Countries
300 Visionare
301 Impact of Design on Eductation
302 Design Blue Print
303 Flat Icons
304 Visionare
305 Landing Pages
306 Ingographic Vector Graphs and Elements
307 The problem with Projects
308 Bicycle
309 Infographics Vector Superset
310 The way to Personal Branding
311 Mobile App Smart Phood
312 Rethink Your Website
313 the user experience design process
314 Website Do's and Don'ts
315 Unboxing the iPAD Data
316 IOS app designer guide to cool design and developer love
317 Evolution of Batman
318 Intelligence by Variety
319 The How to Guide to Responsive Email Design
320 Social Media vs Tranditional Media
321 Moon Galaxy
322 Works about contact design company profile
323 The importance of Color Choice in Marketing
324 Responsive Website Design
325 Communication Patterns Around the World
326 Nobel Lureates

327	Web designer profile
328	Body Language in Business
329	What music should you listen to on the job
330	Start up Style
331	Kitchen Cheat Sheet
332	Creative Commons bloggers
333	The history of Web Design
334	Company Profile
335	Western Typefaces
336	Social Media Design Blueprint
337	Forms of Advertising
338	Infographic Definition
339	Creative Routines
340	Modern Business infographics
341	Designer Profile
342	Apple vs Microsoft
343	Art of Mixing Type Faces
344	Packaging Development
345	Labour Market
346	Network Visualization
347	Repsonsive Design
348	The Chematic of Structure
349	Generation Z Marketing
350	Customer Journey Map
351	18 Rules Using Text
352	Nuclear Bomb Power
353	Infographic Elements Bundle
354	Visual Design Trends
355	Serial Entrepreneurs