1 Q&A WIM

1.1 Body Spam: uses, advantages and drawbacks

The Body Spam is a technique used to increase the score of a web page calculated by search engines.
It consists in inserting the keywords directly in the body of the document.
Some advantages are that it is simple and effective but compromises have to be found due to the TF*IDF tecnique used by search engines to counterbalance the value of each word.
Some disadvantages are that the page content is touched and for this reason keywords has to be put in a proper way in order not to unsatisfy the user. So it is also important to provide a sensible and well-structured page to the user, hiding the keywords inserted in the body.
This can be done by inserting the text in white on a white background or by redirecting the user to the page we want him to view. However, in the last case, it is necessary to do it with a Javascript code, so that the search engines would skip the code because it requires so much time to be analyzed.
Another technique, which is not permitted and it is considered ethically incorrect, is the cloaking. It consists in providing different pages under the same address to the search engines bots.
If it is the crawler, it is shown the page with the keywords.
If it is the users, it is shown the page they want to see.
It is the most effective and the most difficult to detect (an employee should come and check the page in person) but it leads to greater penalties in case of being discovered, such as, for example, long-term suspension from indexing and search results.

1.2 Title spam: uses, advantages and drawbacks

The Title Spam is a technique used to increase the score of a web page calculated by search engines.
It consists in inserting the keywords directly in the title of the document.
Some advantages are that it is simple and effective but compromises have to be found due to the TF*IDF tecnique used by search engines to counterbalance the value of each word.
Some disadvantages are that the page content is touched and for this reason keywords has to be put in a proper way in order not to unsatisfy the user. However it touches less the page content and it is something which is less visible by the user because it is outside the content of the page itself.

1.3 Meta tag spam: uses, advantages and drawbacks

The Meta tag spam is a technique used to increase the score of a web page calculated by search engines.
It consists in entering the keywords in the appropriate meta-data tag. The advantage is that the content of the page is not affected.
The disadvantage is that search engines tend not to give much weight to the words entered there, especially if they are repeated several times because it is an simple and highly used technique.

1.4 Anchor text spam: uses, advantages and drawbacks

The anchor text spam is a technique used to increase the score of a web page calculated by search engines.
It consists in inserting the keywords in the names of the links to other pages. In fact the anchors are part of the body of a page but are treated separately.
Here some advantages. Keywords become very visible because links are with different colours and underlined.
Search engines usually tend to give a high weight to the terms inserted in the anchors.
Furthermore, the keywords inserted in the anchors are usually also added to the target pages without being subjected to penalty filters. This happens because an anchor is supposed to give an idea of the page it points to.
A disadvantage can be that it can touch the real content of the pages, so it is important not to put unrelated content in order not to get the user angry.

1.5 URL spam: uses, advantages and drawbacks

The URL spam is a technique used to increase the score of a web page calculated by search engines.
It consists in inserting the keywords directly in the web address of the page. In fact search engines index and use the URLs itself to calculate the scores of the pages, giving bonuses similar to those of anchortext spam.
The advantage is that it does not affect the content of the page and is usually used in combination with other term spam techniques.
Even with this technique it is necessary to act rationally because as disadvantage there is the fact that if you insert the same word multiple times you will be penalized by TF-IDF.

1.6 Repetition technique: uses, advantages and drawbacks

Repetition is a technique used to increase the keyword score on a web page.
It consists in repeating one or a few keywords several times.
In this way it generates the advantage of increasing the relevance of the page with respect to a single one or to a low number of keywords.
There are also some disadvantages. Because it is a technique that is very easily detectable by crawlers, it is necessary, if you do not want to be penalised, to pay particular attention to countermeasures. Another thing to take into consideration is the TF*IDF in order not to be penalised because some keywords are used so many times.

1.7 Dumping technique: uses, advantages and drawbacks

The Dumping is a technique used to increase the score of a web page calculated by search engines.
It consists in inserting a large number of rare terms, even if they are not related to the content of the page. Here some advantages. The page will be relevant for many different terms which will have only a few other relevant pages because they are considered rare.
But there are also disadvantages. Inserting terms not relevant to the content on the page makes more difficult keeping user, even if it facilitates access to the site. This because what our page offers will be probably different from what they were looking for, leading also loss of trust by the user.

1.8 Weaving technique: uses, advantages and drawbacks

The Weaving is a technique used to increase the score of a web page calculated by search engines.
It consists in copying text from other web pages and inserting keywords in random positions. This technique works best if the topic covered by the copied text is rare or for which there are only a few relevant pages.
There are also some other advantages. It is also used to utilyze better the keywords entered, so that they can be repeated several times, reducing the possibility of being penalised by TF-IDF.
It is an automatic way for making interesting content and attracting more the users.
As disadvantage, it affects the content of the page, so it is important not to put unrelated content in order not to get the user angry.

1.9 Stitching technique: uses, advantages and drawbacks

The Stitching is a technique used to increase the score of a web page calculated by search engines.
It consists in paste&copying from different web sources and then assembling everything with the aim of quickly obtaining relevant content.
Here there are some advantages. This technique is useful for quickly populating a site and for making the page have the possibility of be shown in the search results for each of the topics covered in the copied texts. Another benefit of this technique is that many search engines reward sites with more pages.

1.10 Broadening technique: uses, advantages and drawbacks

The Broadening is a technique used to increase the score of a web page calculated by search engines.
It consists in entering synonyms and related phrases in addition to entering the chosen keywords, also. Here some advantages. It is seen very positively by search engines. In particular this helps to cover more user queries because they are not always very precise. There can be also some extra bonuses if the keywords are similars to each others.

As disadvantage, it may be a good thing to control the penalisation made by FT-IDF. In fact, choosing different words can also lead to a decrease of ranking.

1.11 Cloaking: uses, advantages and drawbacks

The Cloaking is a technique used to increase the score of a web page calculated by search engines.
It consists consists in providing different pages under the same address to the search engines bots.
If it is the crawler, it is shown the page with the keywords.
If it is the users, it is shown the page they want to see.
The advantage is that the page is more visible and with an higher rank. It is also the most effective and the most difficult to detect (an employee should come and check the page in person). As disadvantage, it leads to great penalties in case of being discovered, such as, for example, long-term suspension from indexing and search results.

1.12 Describe the LOD classification: its features, uses and potential advantages and drawbacks

Linked Data are a generalization of the semantic web and they are used to create knowledge graphs.
LOD, linked open data, are linked data that can be used free. It is very important to facilitate their creation and management because they can produce some advantage such as making new knowledge and facilitate innovation, offering the possibility to produce better services and applications.
LOD are classified from 1 to 5 stars:

1. The information is available on the web, in any format, under an open license (free to use).
   (example: Images)

2. The same as above with data in a machine-readable structured format.
   (example: Excel)

3. The same as above and using a non-proprietary structured format.
   (example: CSV)

4. The same as above and in semantic web format, therefore URI identifiers are used so that it is possible to point to a single piece of data.
   (example: RDF, OWL)

5. The same as above and with data are linked to other sources to provide context
   (example: Graph)

Another advantage is that, starting from 3-star data, it is possible to reach 4 or 5 stars by giving it a semantic format.
This operation is called a lifting operation and there are also tools to do this. This is done by joining knowledge graphs and using similarity measures in order to see which graphs can be connected.
The opposite operation is called the lowering operation.

Both operations are called mashup, which consist in mixing RDF with structured data.
The main problem of LOD is that it is necessary to use big data analysis algorithms for processing these data. It is also a computational expensive process and the user does not always feel to be able to interface with a database to extract the data he needs.
Concrete examples of LOD are DBpedia and schema.org.

1.13 RDF and RDFS: their features, uses, advantages and drawbacks.

RDF (Resource Description Framework) is a technology that allows to structure information and remove ambiguities.
It provides data semantics and allows machines to understand the information present on the web.
RDF is written with N-triplets instead of XML because of the existing dialects of XML which make information aggregation impossible. This technology allows to describe metadata, relationships and concepts ensuring interoperability between them (RDF advantage).
The information description is based on the basic grammar composed by triple subject, predicate, direct object and each of these three elements can contain strings or URIs. This structure can be visualized as a graph called knowledge graph and the RDF allows the aggregation of multiple knowledge graphs by linking resources identified by the same URI (RDF advantage).
However, the objects must also be classified ensuring a minimum computational cost and this is allowed by the RDF Schema.
The RDFS is a schema with an information structure which is made of classes, sub-classes and individuals for objects, while is made of properties, sub-properties, domains and intervals (range) for verbs. This implies more power for integrity checks and deductions (RDFS advantage).
The RDF has established basic level semantic rules and so the RDF-Schema allows to define ontologies to categorize information taxonomically (RDFS advantage). This is useful and gives more power to face the URI variants problem (one concept can be expressed differently). So RDFS is used in the semantic web and in various applications such as linked data and ontologies.
The main advantages of RDF is that it is well specified and allows data to be decentralized and distributed, so that anyone can create a vocabulary or publish data about other resources. Moreover knowledge graph is conceptually simple to understand and analyze.
Disadvantages of RDF is that it is very abstract and verbose, so it is difficult to write or read manually. Moreover programming RDFS requires a knowledge of basic details such as what is a URI and what is a triple.
Two examples of RDF vocabularies are Dublin Core (DC) and Friend Of A Friend (FOAF).

1.14 Dublin Core: its uses, advantages and drawbacks

The Dublin Core is one of the first attempts to structure the web in a semantic way.
It has got 15 basic and optional metadata elements have been defined. They can be repeatable and independent from the domain in which they are used (Title, Author, Subject, ID, Source, Date, Language, etc...).
An advantage can be that it allows to define the fundamental and essential basic properties for the entire semantic data structure. It also enables interoperability between different systems and platforms and supports the discovery and reuse of digital resources.
A disadvantage can be that it is too generic to adequately describe specific resources.

1.15 FOAF: its uses, advantages and drawbacks

FOAF (Friend of a Friend) is a semantic web vocabulary used to describe social networks and relationships between individuals and organizations. FOAF is expressed in RDF (Resource Description Framework) and enables users to create machine-readable descriptions of their social connections, interests, and personal information.
FOAF has applications in social media, personal home pages and research projects. It allows users to connect and share information with each other and provides a standardized format for describing people, organizations, and their relationships.
Strengths of FOAF include:

- Decentralized structure, which enables users to control and manage their own data
- Easy integration with other web technologies and vocabularies
- Enables data to be linked and shared across multiple websites and platforms
- Provides a machine-readable format for storing and retrieving social data

Weaknesses of FOAF include:

- Low adoption rate among websites and users
- Limited ability to describe more complex relationships and attributes
- Lack of standardization in the use of FOAF vocabulary and the data it represents
- Security and privacy concerns with sharing personal information in a machine-readable format.

1.16 SPARQL: its uses, advantages and drawbacks

SPARQL (SPARQL Protocol And RDF Query Language) is a language that is used to make queries in the semantic web.
It follows the same syntax of SQL, so it has a limited power but with the advantage of being decidable with a computational complexity PSPACE.
SPARQL works on graph structures based on pattern matching between triplets and a query has the following structure:

- PREFIX
- SELECT
- FROM
- WHERE
- ORDERED BY

An interesting aspect of this language is the possibility of searching with optional data that could be null or not present. In the semantic web, in fact, it is easy to have partial information and for this reason the OPTIONAL operator has been inserted in SPARQL. In this way, if in the query there are parameters in OPTIONAL and some data associated with these parameters are absent, the query will not give an error.
A disadvantage of SPARQL is found in particularly complex queries, as the computational cost is noticeable.

1.17 OWL: its uses, advantages and drawbacks

RDF allows automatic aggregation of knowledge through URIs, but there can be issues of ambiguity. In fact, the same resource could be present on the web under different URIs. This problem is called the URI variants problem and is one of the main reasons that an additional ontological layer has been created, the OWL language (Web Ontology Language).
OWL is a W3C language for ontology, it is used to connect vocabularies by defining relationships between them and allowing to indicate when two classes or properties are the same and when they are different.
Here some advantages. It is an enriched schema with more properties and reduce the URIs problem. Computationally speaking, OWL is mostly used in the average case and so it can be considered polynomial. As disadvantage, it is not possible to have both high expressivness and decidability, but a compromise has to be adopted.
Both expressiveness and logical decidability can be guaranteed by the existance of three different subsets of OWL:

- OWL Lite: with limited but decidible expressiveness with a computational complexity SHIFT;
- OWL DL: less limited in expressiveness, always decidable but with a computational complexity SHOIN;
- OWL Full: exploits more advanced logics, impling high expressiveness, but is not decidable.

1.18 Describe at least 1 optimal structure of an alliance to optimize PageRank

In order to boost the hypertextual score of a page people can use the spam farm technique which consists in the creation of structures which collect flows from other websites and then pass it to the target page.
So an optimal structure can be constructed with some pages which are bidirectional linked to the target one. It mantains a very important property called reachability. It consists in the fact that every page must be reachable by the search engine. In fact if a page is not reached by the serch engine, it is like it doesn't exist.
Based on this, there are 4 different main possibilities to create an alliance:

- Deep alliance: given 2 target pages owned by different entities the alliance consists in having 2 target which are double-linked with its pages and the allied pages.
  In this way the ranking score is the average of the 2 target pages. This implies more robustness and stability.

- Superficial alliance: given 2 target pages owned by different entities the alliance consists in sharing all the unidirectional links of the 2 pages double-linking the 2 targets, creating a vortex of flow.
  In this way it is more mantainable because there is only 1 double link between the 2 targets. So the number of links are minimized.
  The score for each target is an increment proportional to the other's allied pages and this implies that the score of each target $>$ the maximum of the original target score.

- Ring: given 3 or more target pages owned by different entities the alliance consists in sharing all the unidirectional links of the creating a unidirectional vortex of flow betwwen the targets.
  This ensures scalability and he score for each target is an increment proportional to the other's allied pages and this implies that the score of each target $>$ the maximum of the original target score.

- Complete core: given 3 or more target pages owned by different entities the alliance consists in sharing all the unidirectional links of the creating a bidirectional vortex of flow betwwen the targets.
  This ensures scalability and the score for each target is an increment proportional to the other's allied pages and this implies that the score of each target $>$ the maximum of the original target score.