From data quality to language quality: the challenges of the fitness for use principle

Workshop
AUTOMATED TEXT GENERATION:
THEORY, METHODS, EXAMPLES

Laurence Dierickx
06-10-2022
TU Dresden

ULB UNIVERSITÉ LIBRE DE BRUXELLES

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Quality principles in news automation

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

What's data quality?

Dimensions of data quality = insufficient

Constructivist approach = evolutive

- Empirical data resulting from observations
- Application domains
- Concepts
- Norms
- User needs

Image source: https://bluwaveanalytics.com/growth/data-quality/

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

What's language quality in news automation?

Multidimensional approach

Related to a given application domain

Lexical, syntaxis, semantics aspects

Evaluation of the system (NLP): metrics but AT or TS ≠ NLG

In journalism: users' perception (well written, objective, pleasant…)

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

What's information quality?

Quality means the degree or level of overall excellence of a news story. It signifies an evaluation of the goodness of a communication message.
Shyam Sundar, 1998

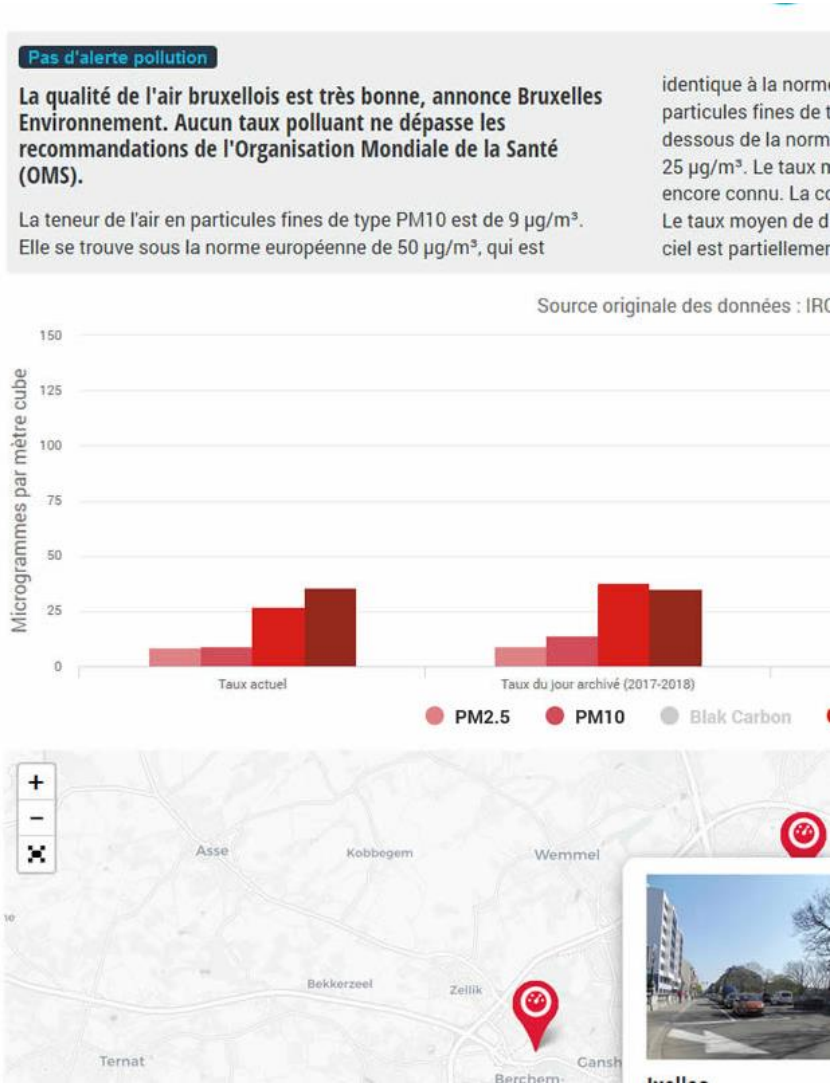
A general, but somewhat vague notion of 'this was a good story'
Christer Clerwall, 2014

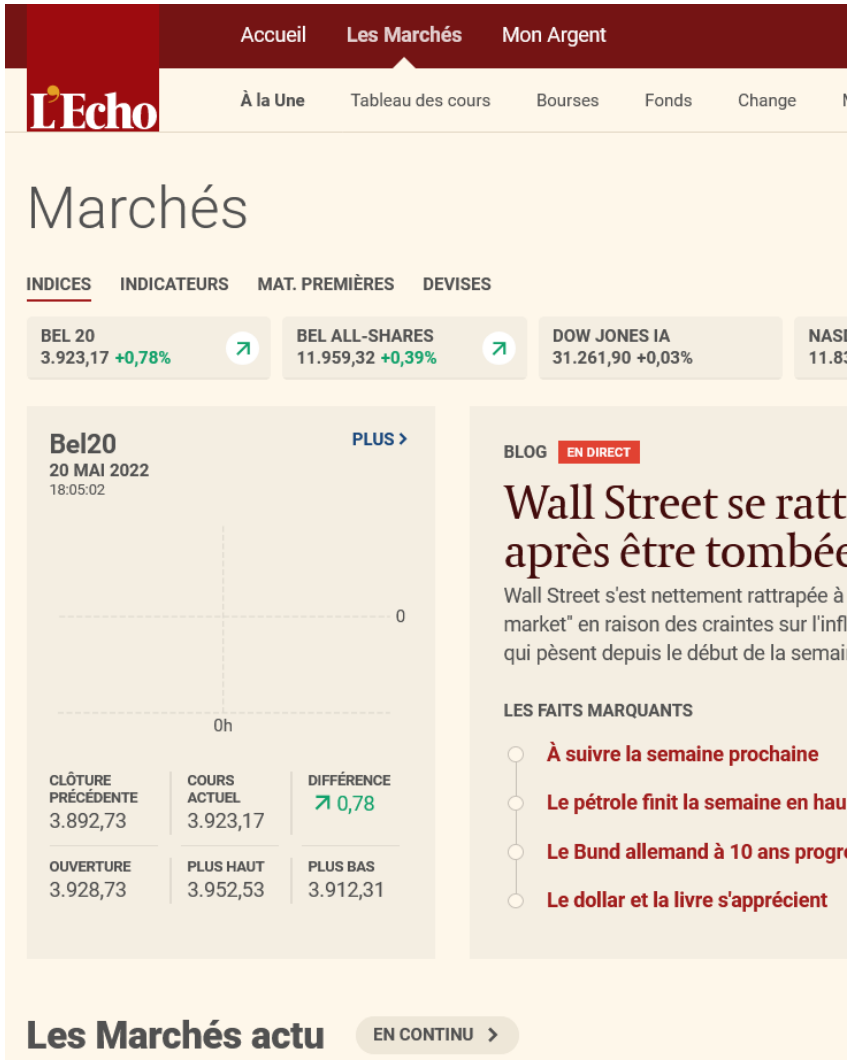
News automation: accuracy, comprehensibility, timeliness, reliability and validity
Nick Diakopoulos, 2019

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

News automation to answer journalists' needs





ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Two research fields / Action research strategy

Bxl'air bot/Alter Echos	Quotebot/L'Echo
Associative organisation	Commercial organisation
Limited proficiency technology	Technology = tools
Limited data literacy	Stock market news = numbers
Monthly, 1k readers	Daily, > 50k readers
6 journalists concerned / 1 year	6 journalists concerned / 2 years
2 journalists involved	3 journalists involved
No budget (web server)	Budget + 211.000 € (DNI)

Dierickx, L. (2020). The social construction of news automation and the user experience. *Brazilian journalism research*, *16*(3), 432-457.

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Assessing quality: users' requirements

Reporting on air pollution in Brussels (local news)

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

A conceptual framework to assess data quality

Technical challenge		
Axis	**Assessment**	
Encoding	No encoding problem No HTML overload	No duplicate data
Normative	Use of standards (date, geolocation,…)	
Semiotic	No missing value No orthographical incoherence	Explicit labelling
Documentary	Unique identifier Available metadata	Conformity metadata/data-set Terms of use

Journalistic challenge		
Contextual	Primary Source (authentic) Appropriate amount of data	Completeness (no missing values) Relevance
Intrinsic	Accuracy (syntactic correctness) Precision (no anomalies observed in values)	Correctness (last update mentioned)

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

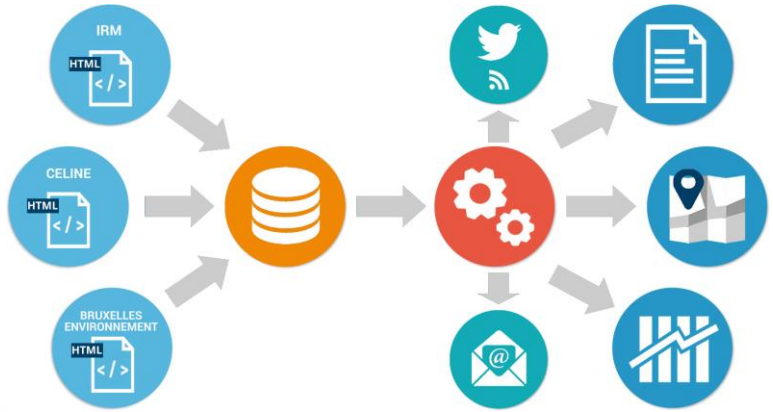
A conceptual framework to assess data quality

Axis	Questions
Source	Is the data provider the producer and/or the authentic source? In the case of the data provider is not the original producer and/or the authentic source, what is the nature of its relationship with the original producer of the data and/or the authentic source? Are the data provider, the data producer, and the authentic source of data trustworthy?
Access	Are data freely accessible? Are they licensed for free reuse? Are they available in a structured format?
Documentation	Are data documented by metadata or any other type of information that allows to understand the database's structure and/or removing any ambiguities in the data labelling? Is any expertise provided to understand the values ? Are contextual elements provided?
Automation	Are data provided in a free and usable format? Do the data values meet the standards? Are the values accurate? Is the dataset complete and up-to-date?
Journalistic relevance	Do data have journalistic added value? How does the data processing make sense?

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Bulletin du [date] - [alerte] - [heure]. [indice][valeur] [Bruxelles], [source des données]. [PM 10] [valeur]. [comparaison] [valeur] [recommandation OMS]. [comparaison] [valeur] [norme européenne]. [comparaison] [valeur] [Belgique]. [PM 2.5] [valeur]. [comparaison] [recommandation OMS]. [comparaison] [Belgique]. [Black carbon] [valeur]. [comparaison] [Belgique]. [O3] [valeur]. [comparaison] [OMS]. [NO2] [valeur]. [comparaison] [Belgique]. [type de temps]. [température].



Bulletin du 15 juillet 2017

Pas d'alerte pollution

14h30. L'indice de la qualité de l'air est très bon en région bruxelloise, indique Bruxelles Environnement. La teneur de l'air en particules fines de type PM10 est de 13 µg/m³. Elle est sous la norme européenne de 50 µg/m³, laquelle est identique à la norme recommandée par l'Organisation mondiale de la santé (OMS). Ce taux se trouve sous le taux moyen constaté à l'échelle nationale (15 µg/m³). Le taux moyen de particules fines de type PM2.5 est de 9 µg/m³.

Ce taux se trouve en-dessous de la norme recommandée par l'OMS. Il est plus élevé que le taux moyen observé à l'échelle nationale. Le tau moyen de particules fines de type black carbon (carbone suie) est de 0,8 µg/m³. La concentration d'ozone dans l'air est de 58 µg/m Le taux moyen de dioxyde d'azote est de 15,9 µg/m³. La couverture du ciel est très nuageuse, pour une température de 19,9 degrés.

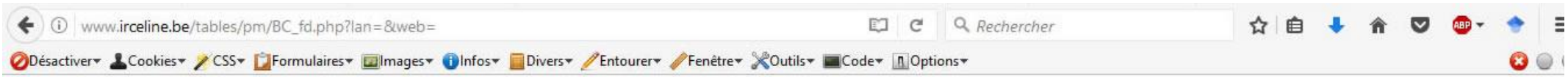
Ce texte a été généré de manière automatique à partir de données publiques extraites en temps réel.

Texts can be reproduced in the sole context of measuring atmospheric pollutant levels.

previous 14 days

error error error error error error error error error error error error error error

code		05/02	06/02	07/02	08/02	09/02	10/02	11/02	12/02	13/02	14/02	15/02	16/02	17/02	18/02
41B004	Brussel (Sint-Katelijne)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41B006	Brussel (EU Parlement)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41B011	Sint-Agatha-Berchem	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41N043	Voorhaven (Haren)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41R001	Sint-Jans-Molenbeek	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41R012	Ukkel	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41WOL1	Sint-Lambrechts-Woluwe	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA



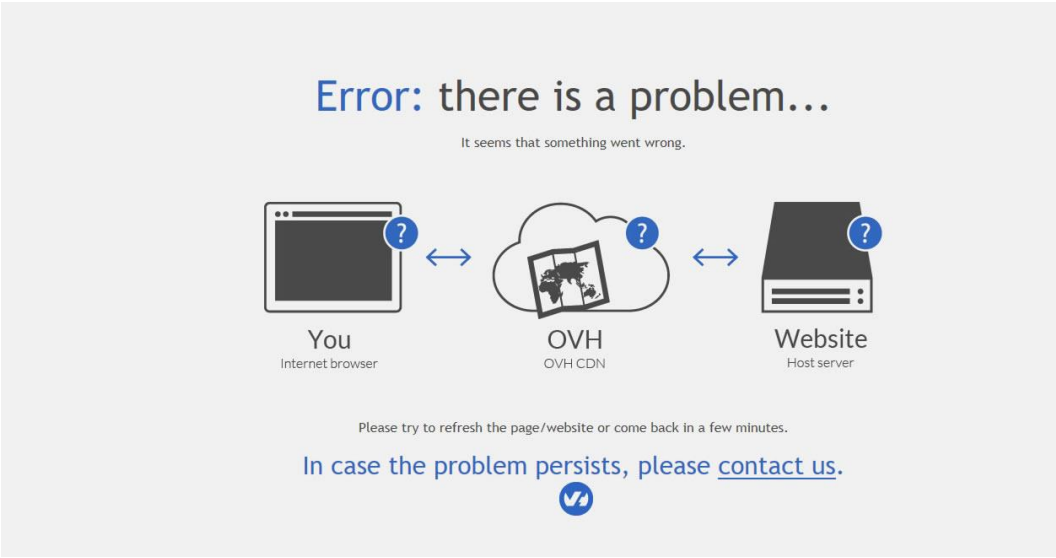
previous 14 days

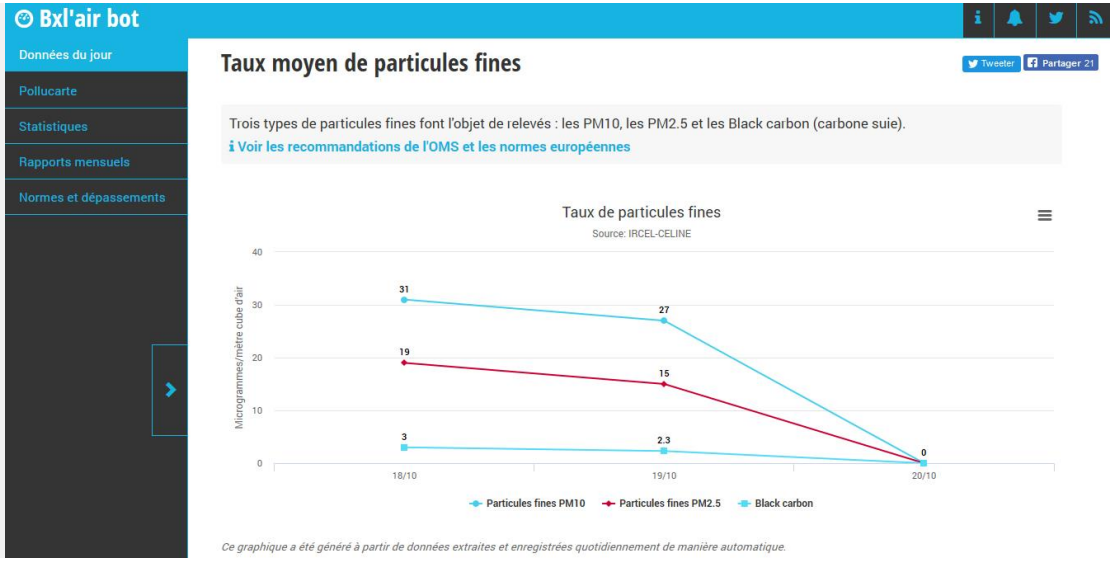
Black Carbon (BC) : Daily mean concentrations (00h00 till 24h00 GMT), last 14 days

code	city	23/05	24/05	25/05	26/05	27/05	28/05	29/05	30/05	31/05	01/06	02/06	03/06	04/06	05/06
41N043	Voorhaven (Haren)	1.8	1.8	0.7	1.1	1.3	0.8	2.0	1.7	1.5	1.3	2.3	1.8	0.9	NA
41R001	Sint-Jans-Molenbeek	NA	NA	NA	NA	0.8	0.8	1.5	0.7	0.8	1.5	1.3	0.9	0.5	NA
41R002	Elsene	1.9	2.0	0.7	0.8	1.5	1.2	2.7	1.8	1.7	1.3	2.4	2.1	1.4	NA
41R012	Ukkel	0.4	0.4	-2.8	0.3	-1.2	0.4	0.7	0.3	0.4	0.5	0.7	0.4	-1.4	NA

ULB

UNIVERSITÉ LIBRE DE BRUXELLES





Bulletin du 3 juillet 2017

Tweeter Partager 12

Pas d'alerte pollution

10h33. Plusieurs données ne sont pas disponibles aujourd'hui, l'indice de qualité de l'air y compris. La teneur moyenne en black carbon (carbone suie) est de 1,8 µg/m³. La concentration d'ozone dans l'air est de 39 µg/m³. Le taux moyen de dioxyde d'azote est de 38,6 µg/m³. Le ciel est peu nuageux, pour une température de 18,1 degrés.

Ce texte a été généré de manière automatique à partir de données extraites en temps réel.

Un an avec un robot

Pendant douze mois, *Alter Échos* a accueilli le Bxl'air bot à la rédaction. Cette application de datajournalisme nous a aidés à enregistrer, compiler et compter des données sur la qualité de l'air. Le résultat ? Pour vous, de l'info inédite sur la pollution à Bruxelles. Pour nous, une première expérience humano-robotique.

PAR CÉLINE GAUTIER

Il a mis le pied dans la porte de la rédaction en mars 2017 sans qu'on l'ait vraiment invité... Le Bxl'air bot est arrivé avec l'enthousiasme de sa conceptrice, la journaliste et développeuse Laurence Dierickx, dans le cadre de son doctorat en information et communication à l'ULB. Son idée : tester, pendant un an et pour la première fois en Belgique, l'immersion d'un robot d'information (ou « newsbot ») dans un média.

Le Bxl'air bot ne prend pas beaucoup de place et ne sert pas le café. Il s'agit d'une simple application qui a fait son nid sur notre site internet. Du 1er avril 2017 au 31 mars 2018, ce « baby bot » a audité, chaque jour, la qualité de l'air dans la capitale, sur la base des données publiées par CELINE, la Cellule interrégionale de l'environnement. Il a produit, minutieusement, son petit rapport quotidien, encore disponible sur http://bxlairbot.be/

Jusqu'ici, les données brutes fournies par CELINE étaient surtout utilisées par les autorités régionales bruxelloises pour communiquer au grand public un indice de la qualité de l'air (voir qualitedelair.brussels/) et pour rendre des comptes à l'Europe sur les niveaux de pollution. L'intérêt du robot, c'est qu'il peut automatiser des calculs que les journalistes pourraient faire manuellement avec les données de CELINE mais qu'ils n'auraient – pour être honnêtes – jamais le temps et la patience de faire.

DE L'OBJECTIVITÉ DU ROBOT-JOURNALISTE

Tout comme les autorités, qui interprètent les données d'une façon rassurante pour le public (voir plus loin notre graphique « La communication des autorités »), le robot assume la part de subjectivité que comporte toute interprétation d'informations. « *Le robot a une approche journalistique*, défend Laurence Dierickx. *Il tient compte de la santé publique.* » En clair, cela signifie que, pour chaque résultat publié, Bxl'air bot le met en relation avec les normes – non pas celles fixées par l'Europe et qui tiennent compte de réalités économiques (et du poids de la bagnole dans notre économie) mais celles proposées par l'Organisation mondiale de la santé (OMS) qui visent, plutôt, à protéger nos cœurs et nos poumons.

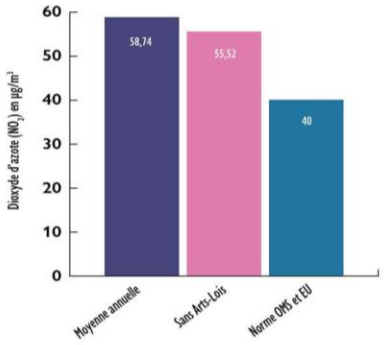
Par exemple, quand CELINE enregistre une moyenne de 27 microgrammes par mètre cube (µg/m³) de particules fines PM10, l'Europe inspire à fond (elle tolère une moyenne annuelle de 40 µg/m³), alors que l'OMS se tord de douleur (elle recommande de ne pas dépasser 20 µg/m³). C'est donc en se basant sur des normes strictes et plus respectueuses de notre santé que le robot conclut, après un an d'enregistrement : « *En moyenne, pour l'ensemble de la Région, les recommandations de l'OMS ont été dépassées deux fois en ce qui concerne le taux de particules fines de type PM10, et 20 fois en ce qui concerne celui des PM 2,5. La recommandation de l'OMS relative au taux d'ozone a été dépassée à 28 reprises.* » Pour rappel, la mauvaise qualité de l'air serait responsable, selon l'Agence européenne de l'environnement, de 12.000 décès prématurés en Belgique par an (pneumonies, cancers, accidents cardiovasculaires, etc.). Ces chiffres valaient bien un robot.

GRAPHIQUE 1*

BRUXELLES EN INFRACTION

Moyennes annuelles de dioxyde d'azote (NO_2) en µg/m³

CELINE, la Cellule interrégionale de l'environnement qui gère les stations de mesure, enregistre des données à Arts-Loi. Sur ce carrefour très embouteillé, les concentrations de dioxyde d'azote (NO_2), un gaz lié au transport routier et au chauffage, sont les plus élevées de la Région bruxelloise. Ces données sont bien disponibles sur le site de CELINE mais ne sont pas prises en compte dans les moyennes, ni pour communiquer au public l'indice de la qualité de l'air ni pour rendre des comptes à l'Europe sur les niveaux de pollution en Belgique. Argument de la Région : Arts-Loi n'est pas représentatif de Bruxelles car trop pollué (la station de mesure est trop proche des voitures). Cette interprétation a été dénoncée (jusqu'au tribunal), depuis des années, par des militants pour la qualité de l'air, qui estiment au contraire qu'il faut tenir compte des pires situations, comme des meilleures, pour avoir un tableau complet de l'enfumage bruxellois.



Frondeur dans l'âme, notre petit robot nous fournit donc une moyenne annuelle de la concentration en dioxyde d'azote à Bruxelles en tenant compte des mesures d'Arts-Loi. Le résultat obtenu (58,74 µg/m³) est bien plus élevé que la moyenne calculée sans Arts-Loi (55,52). Mais dans les deux cas, on dépasse – et de loin – la limite de 40 µg/m³ fixée tant par l'Organisation mondiale de la santé (OMS) que par l'Union européenne (UE). Si le dioxyde d'azote (NO_2) était un gaz hilarant, à Bruxelles, on serait déjà tous morts de rire.

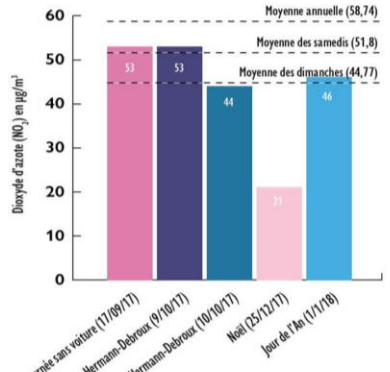
GRAPHIQUE 2

L'EFFET DU DIMANCHE

Dioxyde d'azote (NO_2) en µg/m³

Ce graphique montre l'effet du week-end sur les émissions de dioxyde d'azote. Par rapport aux jours de semaine, la pollution diminue le samedi et baisse encore notablement le dimanche. Ces jours-là, il y a en effet moins de voitures et moins de chauffage dans les bureaux et les collectivités.

Nous avons également repris dans le tableau les résultats enregistrés certains jours de l'année. Pour pouvoir faire de réelles comparaisons, il faudrait tenir compte des conditions climatiques et répéter l'expérience plusieurs années d'affilée. Mais notons, à titre indicatif, que Noël était le jour où nous pouvions respirer le plus à l'aise. Et que les jours de grève ou de fermeture d'un viaduc, les résultats semblent se rapprocher de ceux d'un week-end. Parce qu'en cas de force majeure, on finit par prendre moins sa voiture, privilégier le covoiturage ou travailler de chez soi ? ➔



Les graphiques tiennent compte des données enregistrées par le robot entre le 1er avril 2017 et le 31 mars 2018 sur le territoire de Bruxelles-Capitale.

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Assessing quality: users' requirements

Reporting on stock market news

Échantillon de texte : ouverture à la hausse à la bourse de Bruxelles

La Bourse de Bruxelles était bien orientée dans les premiers échanges, avec un indice Bel20 qui gagnait XXAu sein de l'indicateur, XX valeurs étaient en hausse, XX inchangée(s) et XX en baisse. Les meilleures progressions étaient à mettre sur le compte de/d' Valeurs 1 (+XX%), de/d' Valeurs 2 (+XX%) et de/d' Valeurs 3 (+XX%). À l'autre bout du spectre, on retrouvait Valeurs 4 (-XX%), Valeurs 5 (-XX%) et Valeurs 6 (-XX%). Hors Bel20, on retiendra les bonds de/d' Valeurs 7 (XX%), Valeurs 8 (XX%) et Valeurs 9 (XX%), ou encore les pertes encaissées par Valeurs 10 (XX%), Valeurs 11 (XX%) et Valeurs 12 (XX%).

La Bourse de Bruxelles était bien orientée dans les premiers échanges, avec un indice Bel20 qui gagnait 0,72% à 3.913,72 points. Au sein de l'indicateur, 5 valeurs étaient en hausse, 2 inchangée(s) et 13 en baisse. Les meilleures progressions étaient à mettre sur le compte d'AB InBev (+2%), de Solvay (+1,7%) et d'Umicore (+1%). À l'autre bout du spectre, on retrouvait Aegas (-3%), Proximus (-1,9%) et Cofinimmo (-0,82%). Hors Bel20, on retiendra les gains de Mithra (+14%), d'Aedifica (+3%) et de Sapec (+1,7%). À l'opposé, Montea cédait 2%, Ter Beke 1,3% et Curetis 1,2%).

OU

Emmenée par Valeurs 1 (+XX%), Valeurs 2 (+XX%) et Valeurs 3 (+XX%), la Bourse de Bruxelles progressait dans les premières minutes de la séance, comme le reflétait le gain de XX% de l'indice Bel20. En queue de peloton, Valeurs 4, Valeurs 5 et Valeurs 6 pesaient sur la tendance et rétrogradaient de respectivement XX%, XX% et XX%.

TEST 18c

Source

Wall Street marque le pas dans les premiers échanges. L'indice Dow Jones perd 0,35% à 26.705,25 points, l'indice S&P 500 0,04% à 2.936,76 points et l'indice composite du Nasdaq 0,39% à 7.477,45 points.

Cible

Wall Street présentait des indices majoritairement en baisse à l'ouverture de la séance, l'indice Dow Jones Industrial Average cédait du terrain, à hauteur de 0,35% à 26.705,25 points. L'indice élargi S&P 500, pour sa part, était en légère baisse de 0,04% à 2.936,76 points et l'indice Nasdaq perdait 0,39% à 7.477,45 points.

TEST 28c

Source

Wall Street a clôturé en hausse ce dimanche : le Dow Jones a progressé de 0,14%, tandis que le S&P 500 reculé de 2,15% et le Nasdaq de 0,03%.

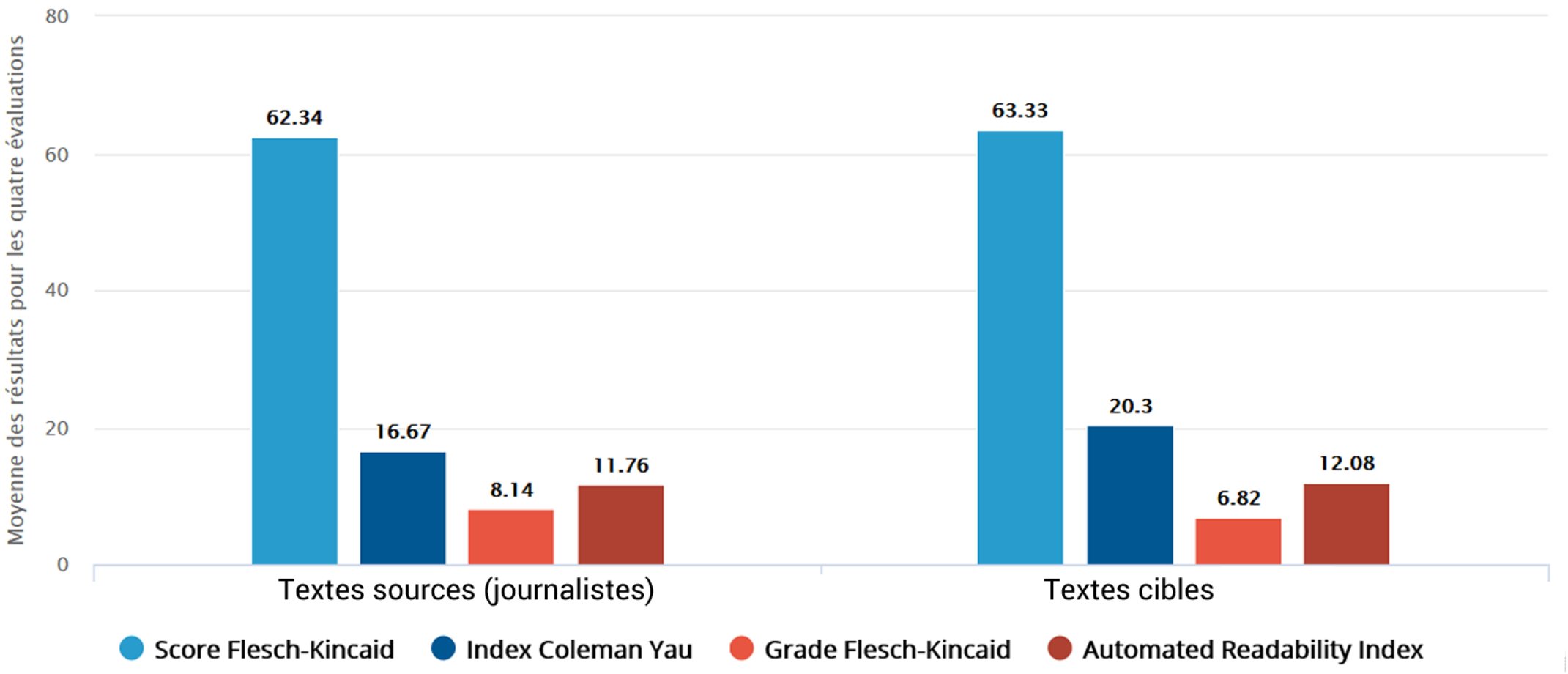
Cible

À la clôture de la séance, l'hésitation dominait à la Bourse de Wall Street, l'indice Dow Jones a perdu 0,14% à 25.762,54 points. Le S&P 500 a gagné 2,15% à 2.809,92 points tandis que l'indice composite du Nasdaq était en légère hausse de 0,03% ce dimanche, à 7.648,02 points.

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Automatic assessments



ULB

UNIVERSITÉ LIBRE DE BRUXELLES

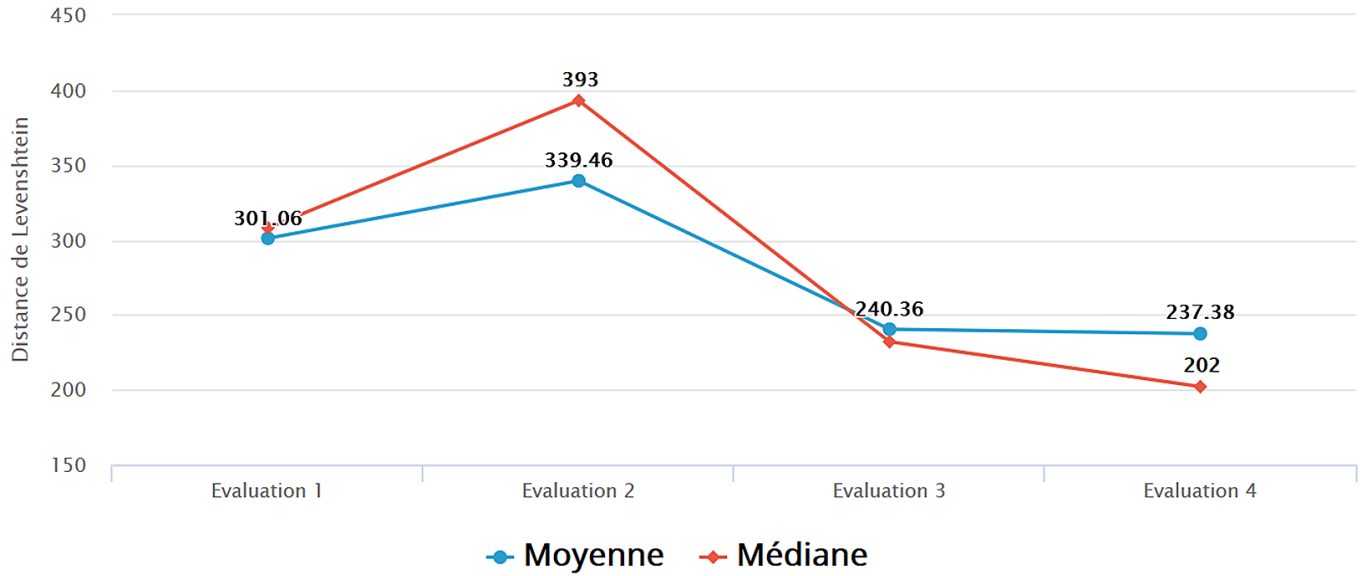
Distance de Levenshtein

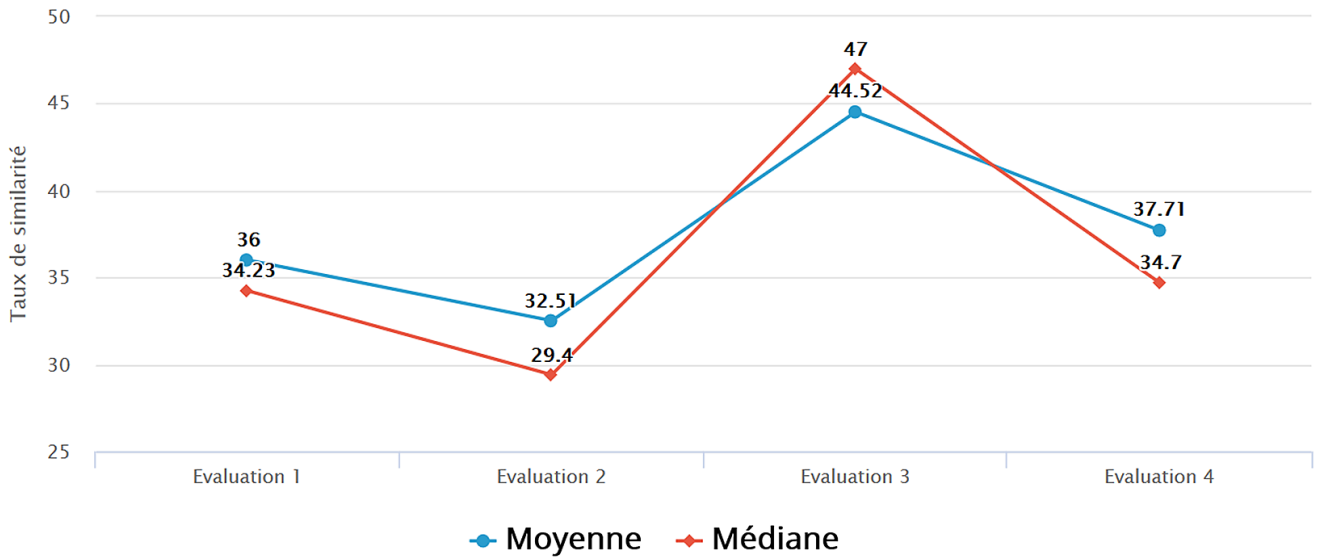
Algorithme en langage de programmation Javascript

```
function distance(a, b) {
	var n = a.length, m = b.length, matrice = [];
	for(var i=-1; i < n; i++) {
		matrice[i]=[];
		matrice[i][-1]=i+1;
	}
	for(var j=-1; j < m; j++) {
		matrice[-1][j]=j+1;
	}
	for(var i=0; i < n; i++) {
		for(var j=0; j < m; j++) {
			var cout = (a.charAt(i) == b.charAt(j))? 0 : 1;
			matrice[i][j] = minimum(1+matrice[i][j-1], 1+matrice[i-1][j], cout+matrice[i-1][j-1]);
		}
	}
	return matrice[n-1][m-1];
}
```

Fonction native de PHP

levenshtein (string $str1 , string $str2) : int





Taux de similarité

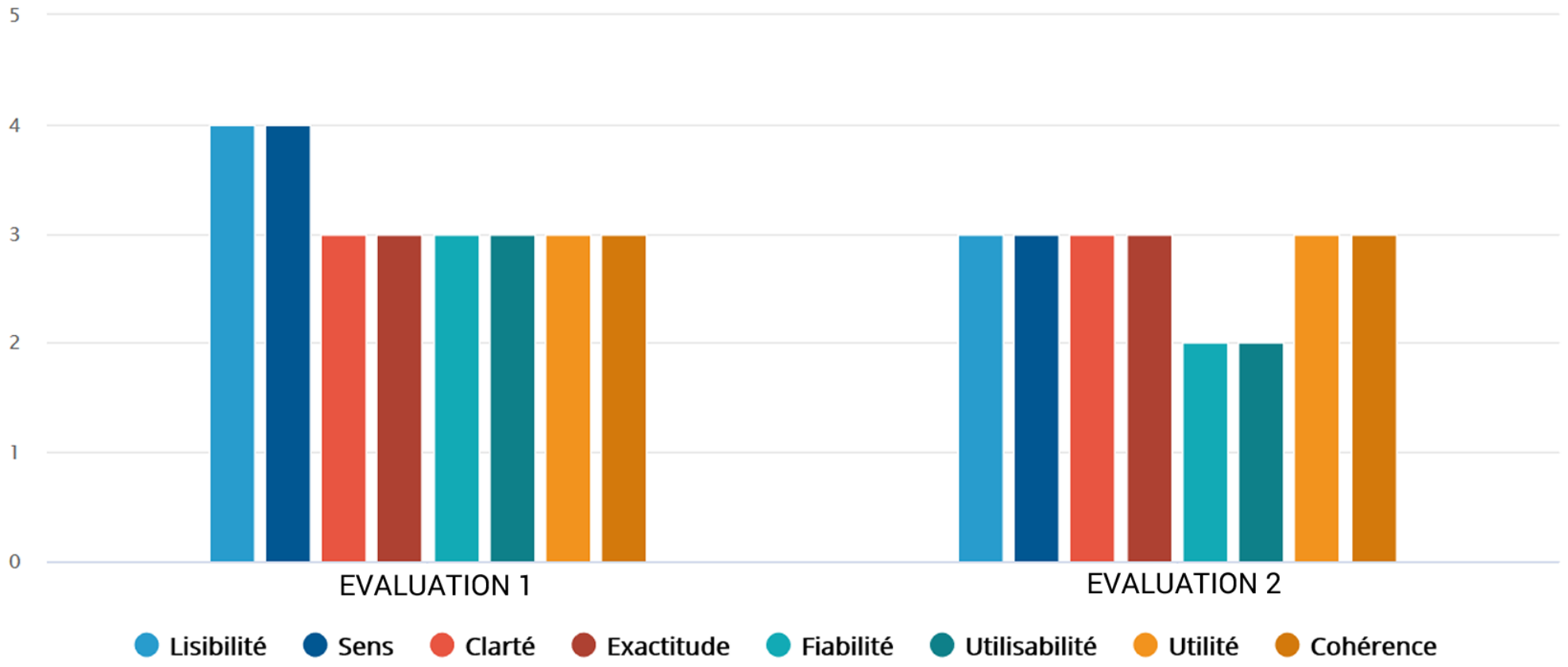
similar_text($source, $cible, $similar); echo round($similar,2). "%" ;

Tools used for the assessments: https://journodev.tech/linguistics

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Human evaluation



ULB

UNIVERSITÉ LIBRE DE BRUXELLES

quotebot lecho

Tous | Images | Maps | Actualités | Vidéos | Plus | Outils

Environ 81 résultats (0,17 secondes)

L'Echo

Wall Street se rattrape à la clôture, après être tombée en "bear ...

... KBC-25,07%;
VGP-26,33%;
Sofina-40,14%. Les chiffres sont susceptibles d'évoluer à la marge. Rédigé p...

Il y a 1 jour

L'Echo

Le Nasdaq chute de près de 5%

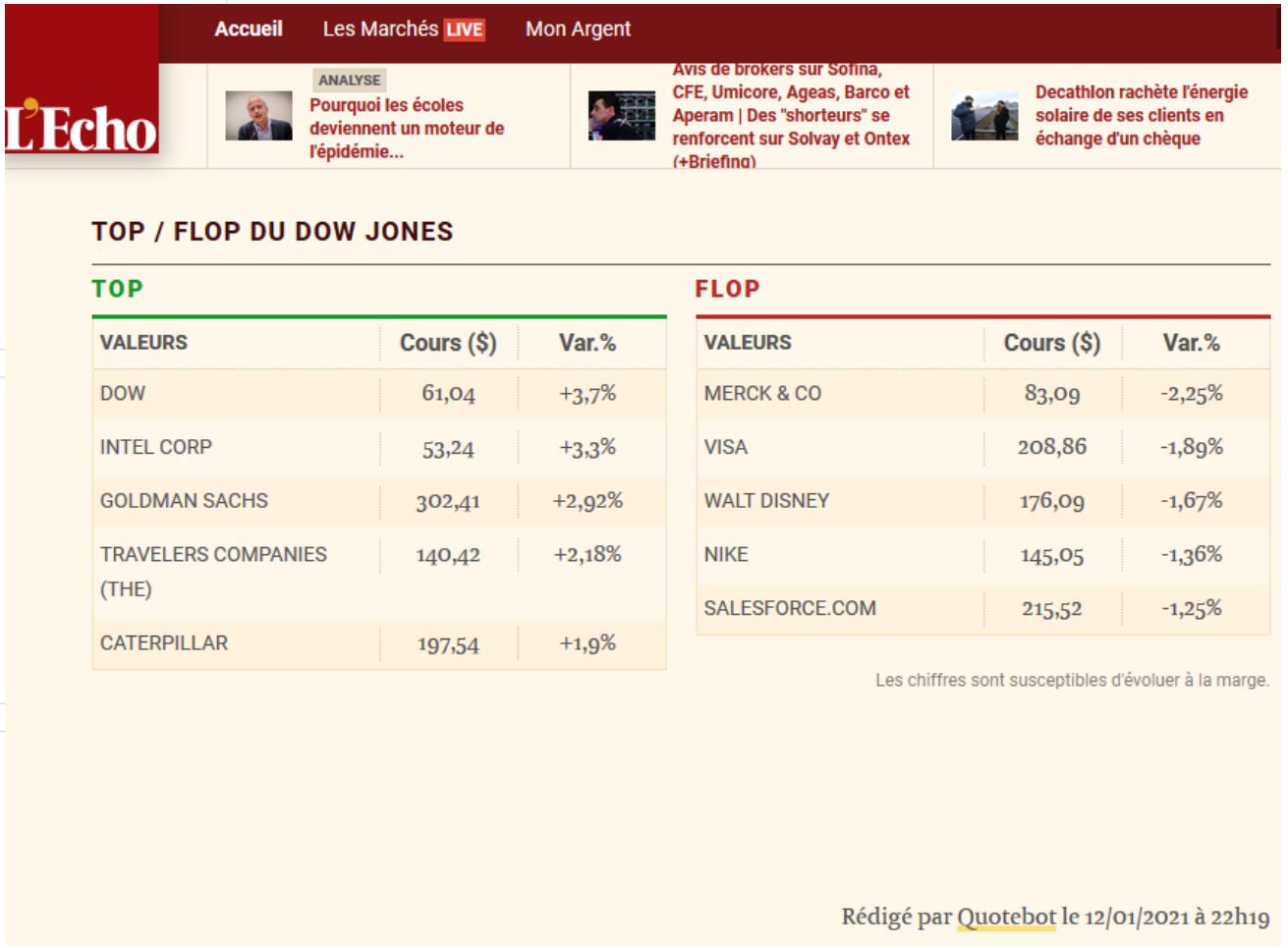
Les chiffres sont susceptibles d'évoluer à la marge. Rédigé par Quotebot le 18/05/2022 à 17h51.

Il y a 3 jours

L'Echo

Le S&P 500 en "bear market"

Les chiffres sont



L'Echo

Accueil | Les Marchés LIVE | Mon Argent

ANALYSE
Pourquoi les écoles deviennent un moteur de l'épidémie...

Avis de brokers sur Sofina, CFE, Umicore, Ageas, Barco et Aperam | Des "shorteurs" se renforcent sur Solvay et Ontex (+Briefing)

Decathlon rachète l'énergie solaire de ses clients en échange d'un chèque

TOP / FLOP DU DOW JONES

TOP

VALEURS	Cours ($)	Var.%
DOW	61,04	+3,7%
INTEL CORP	53,24	+3,3%
GOLDMAN SACHS	302,41	+2,92%
TRAVELERS COMPANIES (THE)	140,42	+2,18%
CATERPILLAR	197,54	+1,9%

FLOP

VALEURS	Cours ($)	Var.%
MERCK & CO	83,09	-2,25%
VISA	208,86	-1,89%
WALT DISNEY	176,09	-1,67%
NIKE	145,05	-1,36%
SALESFORCE.COM	215,52	-1,25%

Les chiffres sont susceptibles d'évoluer à la marge.

Rédigé par Quotebot le 12/01/2021 à 22h19

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Conclusion and way forward

- Data quality = angular stone + expertise of the application domain
- Metrics allow to detect and prevent errors (data),
  to objectivate human perception even if no correlation (language)
- Human judgements can contradict metrics, but new metrics appeared in the wake of the development of ML systems (BLEURT)
- Challenges of quality evolve but the same preoccupations with ML and NN systems, although specific measures of accuracy etc.
- Need for interdisciplinarity

Thank you for your attention!

https://journodev.tech/dresden061022/

ULB UNIVERSITÉ LIBRE DE BRUXELLES