Web Data Processing Systems 2022/2023

Jacopo Urbani

jacopo@cs.vu.nl

VU VRIJE UNIVERSITEIT AMSTERDAM

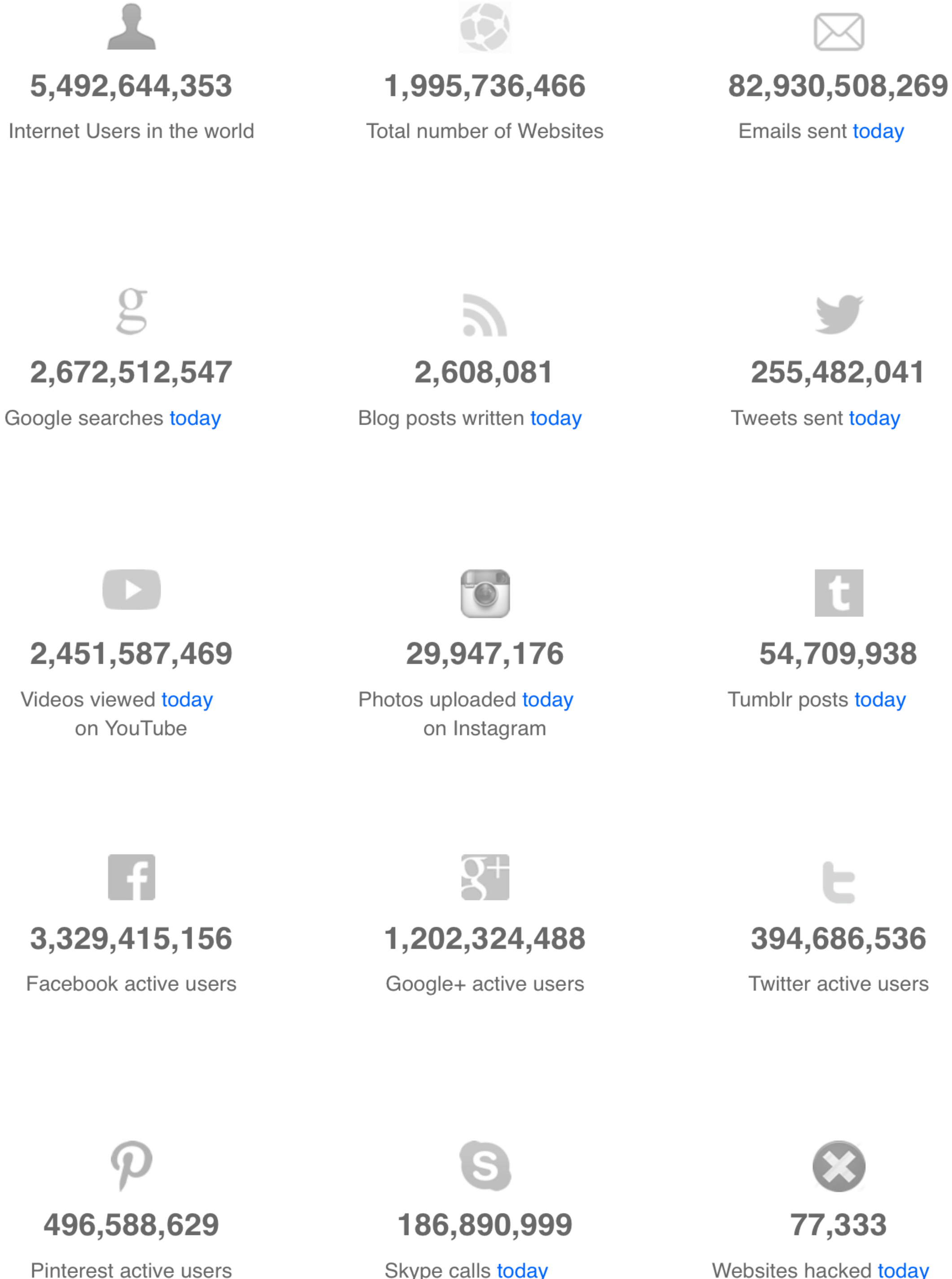
The Web

The largest body of knowledge ever assembled

Applications

- Answering questions
- Recognizing objects
- Store knowledge
- Improve education
- …

VU VRIJE UNIVERSITEIT AMSTERDAM



5,492,644,353
Internet Users in the world

1,995,736,466
Total number of Websites

82,930,508,269
Emails sent today

2,672,512,547
Google searches today

2,608,081
Blog posts written today

255,482,041
Tweets sent today

2,451,587,469
Videos viewed today
on YouTube

29,947,176
Photos uploaded today
on Instagram

54,709,938
Tumblr posts today

3,329,415,156
Facebook active users

1,202,324,488
Google+ active users

394,686,536
Twitter active users

496,588,629
Pinterest active users

186,890,999
Skype calls today

77,333
Websites hacked today

The Web

Ground for many lucrative businesses

- Information search (Google)
- Social media, networking (Facebook, LinkedIn, Twitter)
- E-commerce (Amazon)
- Access economy (Uber, Airbnb)
- Sharing economy (crowdfunding, Wikipedia)
- etc.

Sandhill road, Palo Alto

This course

In a nutshell

Goals

- **Understand** and **extract** knowledge from Web content
- **Infer** new knowledge from Web content
- **Verify** information on the Web
- **Protect** users’ privacy
- Ensure **fairness** while searching for content

Challenges

- **Size***:* The Web is huge so scalability is paramount
- **Uncertainty***:* Extraction techniques make errors
- **Conflicts:** What to do when there are conflicts in the data?

In a nutshell

Approach

- We will look at the most recent developments in the field (*research articles*)
- A particular emphasis will be given to *systems* (rather than theory)
- We will also experience how to build an extraction systems from Web data

Requirements

- Although this is **NOT** a ML course, you will have to use some ML techniques
- This is a *system-oriented* course. Programming skills are required

Topics
Knowledge bases / NLP
Language models
Knowledge Acquisition
Mining and inference on knowledge graphs / social networks
Reasoning (ontologies, uncertainty)
Fact spotting and checking
Fairness of Ranking and Privacy (notes)

Lectures

- The course is almost the same as last year, but there will be **new** material
- The material will be explained in:
  - **Weekly lectures, on campus**
  - **Pre-recorded videos** (some videos are from the previous years, others will be recorded again)
- The lectures on campus are **NOT meant** to repeat the content of the videos. They are more an possibility to re-visit the slides and for Q&A
- No Q&A means that the lecture finishes earlier
- Some lectures are reserved for group meetings

<u>Live lectures will not be recorded</u>

What should I do to prepare for the exam?

My advice is:

1. Watch the videos, possibly before the lecture
2. Come to the lectures on campus, even if you understood everything in the video
3. If something is still not clear you can:
    a) Contact the teacher
    b) Read the papers mentioned in the slides
    c) Look at old copies of the exams, but keep in mind that this year the exam will be different

Practical Assignment

Goal: Put into practice what you have learned

Part of the grades will be given by a practical assignment, to be done in groups of **four** students

Deadline to submit the assignment: 21/12/2022

Details on the assignments will be on Canvas. For now, please form groups of four students. If you cannot find people, please let me know asap.

Exam

- In the last years, the exam contained only open questions
- This year, the number of students is too high. To keep the grading manageable, the exam will also contain multiple-choice questions
- The exam will take place on 15/12/2022 at 12:15 on campus. **It is not possible to do the exam online (e.g., using proctoring).** You must come to the VU

Final grade

Grading formula: 60% final exam, 40% practical assignment. The grade on the final exam **must be greater or equal than 5.5** to pass the course

For non-CS students

If you do not have a strong background in programming, you might struggle with the practical assignments

Tips

- Try to find a group with someone who knows how to program
- You can still pass the course even if the assignment grade is not high

Final remarks

Important remarks

This is a research-oriented course. We will discuss problems that nobody has solved yet

There is not a single textbook for the course. The studying material consists of research publications, online sources, etc.

Frequently asked questions

- **Question:** Can I pass the course without coming to the campus?
- **Answer:** Yes, because the exam will only contain questions about the material mentioned in the slides. However, you cannot demand that the other members of your group meet on Zoom because you cannot come to the VU

- **Question:** I have no programming experience. Can I pass the course?
- **Answer:** I cannot tell you whether you will be able to pass the course. It largely depends on how fast you can learn the missing skills

Frequently Asked Questions

- **Question:** Can you put the slides (videos) online before the lectures?
- **Answer:** I'll do my best, but I cannot make any promise

- **Question:** How can I contact you?
- **Answer:** Preferably by e-mail, not using Canvas